A Machine Learning Approach for Plagiarism Detection

Al-Sallal, Muna

Award date:
2016

Awarding institution:
Coventry University

A Machine Learning Approach for Plagiarism Detection

By

Muna Alsallal

November 2016

A Machine Learning Approach for Plagiarism Detection

By

Muna AlSallal

November 2016

A thesis submitted in partial fulfilment of the University's requirements for the Degree of Doctor of Philosophy

Abstract

Plagiarism detection is gaining increasing importance due to requirements for integrity in education. The existing research has investigated the problem of plagrarim detection with a varying degree of success. The literature revealed that there are two main methods for detecting plagiarism, namely extrinsic and intrinsic.

This thesis has developed two novel approaches to address both of these methods. Firstly a novel extrinsic method for detecting plagiarism is proposed. The method is based on four well-known techniques namely Bag of Words (BOW), Latent Semantic Analysis (LSA), Stylometry and Support Vector Machines (SVM). The LSA application was fine-tuned to take in the stylometric features (most common words) in order to characterise the document authorship as described in chapter 4. The results revealed that LSA based stylometry has outperformed the traditional LSA application. Support vector machine based algorithms were used to perform the classification procedure in order to predict which author has written a particular book being tested. The proposed method has successfully addressed the limitations of semantic characteristics and identified the document source by assigning the book being tested to the right author in most cases.

Secondly, the intrinsic detection method has relied on the use of the statistical properties of the most common words. LSA was applied in this method to a group of most common words (MCWs) to extract their usage patterns based on the transitivity property of LSA. The feature sets of the intrinsic model were based on the frequency of the most common words, their relative frequencies in series, and the deviation of these frequencies across all books for a particular author.

The Intrinsic method aims to generate a model of author "style" by revealing a set of certain features of authorship. The model's generation procedure focuses on just one author as an attempt to summarise aspects of an author's style in a definitive and clear-cut manner.

The thesis has also proposed a novel experimental methodology for testing the performance of both extrinsic and intrinsic methods for plagiarism detection. This methodology relies upon the CEN (Corpus of English Novels) training dataset, but divides that dataset up into training and test datasets in a novel manner. Both approaches have been evaluated using the well-known leave-one-out-cross-validation method. Results indicated that by integrating deep analysis (LSA) and Stylometric analysis, hidden changes can be identified whether or not a reference collection exists.

Acknowledgment

I greatly thank Allah for giving me the inspiration, patience, time, and the ability to finish this work.

I would like to express my deep and sincere gratitude to my director of study Dr. Rahat Iqbal, for his expert guidance, constructive criticism, continuous support and encouragement throughout my study. I am also grateful to my supervisors, Dr. Saad Amin for his valuable comments and unlimitted support. Special thanks to my supervisor Dr. Vasile Palade, for constructive comments and valuable discussion. I would also like to thank Dr. Anne James for her support and kindness.

I am also indebted to Dr. Mark Elshaw for his kindness and thoughtful constructive comments. I have really appreciated his patience and valuable time. His help and support is very much appreciated.

Gratefull and lots of love also go to Mrs Irene Glendenning for her kind support in all life aspects and general discussion on plagiarism issues. Special thanks to Prof. Ian Dunn for his continuous support.

I want to express my gratitude and deepest appreciation to my best friend Hala Abdulaziz for being in my life. Many thanks are going to all my friends and colleagues at Coventry University in particular colleagues in room EC2.21 for their unlimited support.

I heartily thank my parents for their unconditional love and care; no words can express how much I appreciate what they have done for me.They continuously instilled confedence in me, and provided boundless love. I always make duaa to keep them safe and healthy for ever.

Table of Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
BN	Bayesian Network Machine Learning Algorithm
BOW	Bag of Words
CEN	Corpus of English Novels
CV	Cross-Validation
CW	Content Words
FE	Features Engineering
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive IDF Inverse Term Frequence
LOOCV	Leave-One-Out-Cross-Validation
LR	Likelihood Ratio Metric
LSA	Latent Semantic Analysis

MCW	Most Common Words
MLP	Multilayer Perceptron
MSE	Miss-Classification Error
PCA	Principal Component Analysis
R	A Language and Environment for Statistical Computing and Graphics.
RBF	Radial Basis Function Kernel
RF	Random Forest
SMO	Sequential Minimal Optimization
SVD	Singular Value Decomposition
SVM	Support Vector Machines
SVM-RBF	Support Vector Machines based Radial Basis Function Kernel
SVM-SMO	Support Vector Machines based Sequential Minimal Optimization as Learning Algorithm
TF	Term Frequency
WEKA	Waikato Environment for Knowledge Analysis

Chapter1: Introdution

Plagiarism detection and authorship analysis approaches have a long history of attempts to improve their performance in detecting text misuse and identifying the author of anonymous text. However, despite considerable work in improving such methods, by using different types of features and a wide range of techniques, the performance of these methods is still unsatisfactory in some cases of plagiarism detection. The main goal of this thesis is to investigate those cases and propose a new approach to address plagiarism detection effectively by combining both traditional and machine learning techniques. This chapter focuses on text plagiarism detection challenges and highlights some of the limitations of the existing plagiarism detection tools. The chapter also discusses the research background, the research problem and the motivation for the research. Furthermore, this chapter presents the research aim, objectives, research method and a summary of the research contribution.

The chapter is organised as follows: Section 1.1 presents the background to the research, section 1.2 describes the problem statement as well as the motivation for the research, section 1.3 presents the aim and objectives, section 1.4 explains the research methodology, section 1.5 presents the research scope, section 1.6 presents the research contribution and, finally, section 1.7 outlines the thesis structure.

1.1 Background

The abundance of online information, open resource materials and fast development of networking technologies have encouraged the misuse of information. Moreover, the reduction in price of computing devices and applications has shaped new ways for communication. All these have facilitated

the misuse of materials and created a critical problem of plagiarism across the World Wide Web. According to the Oxford English Dictionary, plagiarism is defined as the action of using knowledge (language, ideas, and findings) without giving proper credit to the author. Murray (2008) stated that plagiarism was the "use or reuse of words or ideas without acknowledgment". Murray tried to differentiate the word "acknowledgement" from "permission" as an indicator that plagiarism was not a legal problem. A group of researchers who preceded Murray had the same point of view regarding plagiarism; they determined that plagiarism was an academic misconduct and not a legal issue (Larkham, and Manns, 2002; Myers, 1998). According to Hannabuss (2001) plagiarism was "the unauthorised use or close imitation of the ideas and language/expression of someone else". Clough (2003) and Piao et al., (2001) have shown different understanding; they argued that the plagiarism concept was still unclear and it was difficult to form a specific definition. Other researchers have given plagiarism a limited definition based on their proposed solution such as in (Sorokina et al., 2006). Sorokina and colleagues defined plagiarism as a series of word n-grams that can be duplicated in different document. Here n-grams meant an adjoining sequence of characters, words, syllables or phonemes or any other text type. They also assumed that this sequence could keep the same characteristics even if the sequence was replaced by synonyms.

In the past, plagiarism detection was carried out manually depending on the investigator's own knowledge. In most cases, the "Déjà vu" sense played an important role in highlighting the suspicious text.

The existing plagiarism detection methods work on identical text matching strategies, without considering the core of the knowledge or how this knowledge is developed. They are inherently limited by their rigid assumptions that plagiarists literally copy and paste whole sentences or paragraphs of text from other authors directly into their documents (zu Eissen and Stein, 2006). Most of the existing plagiarism detection methods are limited and have a number of shortcomings in detecting many types of plagiarism cases (e.g., syntactical or semantics changes) as described later in the chapter (Stappenbelt, B. and

Rowles, 2010; Ramnial, Panchoo and Pudaruth, 2016). The use of the current plagiarism detection tools is limited to indicating plagiarism incidences in order to support human decisions (Alzahrani et al., 2012; Ramnial, Panchoo and Pudaruth, 2016).

The variance between computers and humans in analysing the text has influenced the performance of current detection tools. The meaning of the language of humans is difficult to be understood by computers which greatly affects the human-computer interaction procedures when dealing with the plagiarism problem. Challenges such as semantic changes (i.e. using synonyms to change the vocabulary but keep the meaning) or syntactic manipulations (i.e. re-ordering the words using active voice instead of passive and vice versa) are some of the limitations that current detection algorithms could not deal with (Chong, 2013). Such shortcomings can be overcome using the integration of several effective techniques that can capture the latent associations between words in order to address semantic shortcomings. Kakkonen and Mozgovoy (2010) conducted a comparative study to evaluate the state of the art in plagiarism detection. In their study, they highlighted that the most significant challenge in plagiarism detection field was to identify the text source. They recommended that incorporating authorship detection techniques with the current plagiarism detection approaches could substantially enhance the plagiarism detection performance.

The current literature has split the plagiarism detection methods into two forms: extrinsic and intrinsic (Alzahrani, Salim and Abraham, 2012; Zurini, 2015). Extrinsic plagiarism detection methods rely on comparing the suspicious document or string of text to a body of known, classified documents (Alzahrani, Salim and Abraham, 2012). While these methods perform well to some extent for copy and paste misconduct, the detection assumption was built on a notion that all related information was digitised. A criticism that has been raised against that assumption revolves around the fact that not all sources are digitised (Eissen, B. Stein, and M. Kluig, 2007). Consequently, a new class of plagiarism-detection tools are currently being researched and developed,

termed as intrinsic detection methods. The intrinsic detection methods aim to characterise a writer's style using a history of that writer's existing work (Zechner et al., 2009). These types of detection methods rely on capturing the variations in written text by extracting the syntactic and lexical features. Then, a comparison is performed between the suspicious text and the rest of the author's work in order to capture the variation patterns.

1.2 Problem Statement and Motivation

The ease of sharing online information in this age of digital communication has encouraged the misuse of text and the prevalence of plagiarism. Academic bodies and scientific publishing companies are playing an active role in detecting plagiarism in order to maintain the integrity of academic publications (Glendenning, 2012; Oberreuter et al., 2011; Carroll, 2007).

Redfern and Barnwell (2009) pointed out that many cases of academic work submitted by students contain some level of plagiarised material. Roig (2001) reported that up to 60% of student assignments contain some level of plagiarised material. Few years later, the international centre of academic integrity (ICAI) published that 86% of students were involved in some form of plagiarism (Ramnial, Panchoo and Pudaruth, 2016).

Large publishing companies, such as Springer and Elsevier, claimed that 6% - 23% of articles were refused due to the substantial percentage in the overlapping of information between papers (Sánchez-Vega et, al., 2013; Butler, 2010). Zhang (2010) indicated that 25.8% of submitted articles for publishing in China are considered to have a considerable degree of plagiarism. The above statistics and many more present clearly that the plagiarism problem is growing and being exacerbated in academic work.

A well-known study by Maurer, Kappe, and Zaka (2006) and followed by Maurer and Zaka (2007) provided a comprehensive report on some of the challenges of plagiarism detection systems, such as Turnitin© and Copycatch, and noted down how paraphrasing often renders these tools ineffective. The authors

revealed that existing commercial detection tools were largely unable to cope with synonyms, extensive paraphrasing and cross-lingual plagiarism, resulting in a number of plagiarism cases going undetected. They further recommended the use of an efficient algorithm to extract informative features before running a hybrid algorithm, that can work efficiently on datasets of small documents.

As briefly described in section 1.1, the two most widely recognised methods of plagiarism detection are extrinsic and intrinsic. Yet, the majority of existing detection tools (commercially or freely available) use extrinsic methods and performed identical text-matching (Ramnial, Panchoo and Pudaruth, 2016). Turnitin© and CrossCheck, the leading plagiarism detection software in most academic institutions and publishing firms, are still facing big challenges in detecting linguistic changes such as replacing words by their synonyms (Eisa, Salim and Alzahrani, 2015). They were also criticised owing to their vulnerability for increasing the numbers of false positives (i.e. when cases are detected as plagiarised but in fact are not). As a result they are always in need of human intervention to finalise decisions (Ramnial, Panchoo and Pudaruth, 2016).

A wide range of studies focused on researching and developing methods for effectively combating plagiarism (White, and Joy, 2004; Elhadi, and Al-Tobi, 2008; Alzahrani, Salim and Abraham, 2012). However, there are lack of effective plagiarism detection methods.

Although plagiarism detection tools have been used for a long time in academia, the results always require human intervention to certify whether plagiarism has actually occurred. The interaction process between users and detection tools do not go beyond highlighting the similarity between submitted texts and the repositories maintained by the tools. As a result, users can bypass these detection procedures using appropriate or inappropriate methods. In fact, the tools have failed to achieve their original mission to protect the scientific environment and emphasise ethical issues. On the contrary, users are finding more and more ways to defeat the plagiarism detection algorithms. Furthermore, by targeting the similarities between the words used rather than their meaning they have failed to influence the user understanding of what

plagiarism actually means. Many studies have shown that there is a significant percentage of students and early career researchers who are unable to define the boundaries between plagiarism and original work (Purdy, and Walker, 2013).

There is a significant gap in the literature regarding the research performed in the field of plagiarism detection. Most of the existing detection tools rely on identical string matching techniques ignoring the semantics of the text and the identification of text authorship characteristics, both of which are important issues. Furthermore, they always require close at hand, a complete set of original work to which the text can be compared (i.e. a reference collection).. The human experts can discover plagiarism even if linguistic elements of a piece of text were changed. They also can identify the original source of the manipulated text by comparison against the stylistic attributes of the original author. Several studies were conducted to address these issues with varying degree of success (Ramnial, Panchoo and Pudaruth, 2016; Alzahrani, Salim and Palade, 2015).

The main goal for this research is to address the current detection limitations by engaging intelligent techniques and extending the methods of authorship analysis for the purpose of plagiarism detection. The research intends to develop human like intelligence based approaches to detect plagiarism using both extrinsic and intrinsic methods. The extrinsic method is designed to detect the text semantics and identify the text author when comparison can be made to a reference collection. The intrinsic method needs to identify stylistic variations when no references collection is available for comparison. A set of effective methods are proposed to deal with the plagiarism detection problem as briefly explained above.

The proposed research will address the following main research questions:

1. How effective is the use of latent semantic analysis when combined with stylometry and machine learning approaches for the task of detecting semantic

variations in order to verify the originality of the work of the author when a reference collection is available for comparison?

2. How effective is the use of machine learning approaches based on the most common word frequencies and their derivatives, for the task of detecting variations in style in order to verify the originality of the work of the author, when a reference collection is not available for comparison?

1.3 **Aim and Objectives**

In order to address the research gap and the research questions as described above, the following aim and objectives were set. The main aim of this research is to investigate the existing plagiarism detection techniques and develop an integrated machine learning based approach using latent semantic analysis and stylometry to enhance the performance of these techniques. The aim will be achieved by fulfilling the following objectives:

1) To conduct a thorough literature review of the existing methods, techniques and tools being used for plagiarism detection.

2) To investigate the plagiarism detection types and their limitations.

3) To explore the statistical semantics techniques and their application for enhancing plagiarism detection.

4) To explore the stylometric features of the text sources and the parameters that influence them in order to enhance performance of plagiarism detection techniques.

5) To investigate the role of machine learning approaches for plagiarism detection based on the semantic and stylistic features of the text.

6) To develop human like intelligence based plagiarism detection approaches using machine learning, latent semantic analysis and

stylometry, in order to address the limitations of existing extrinsic and intrinsic detection methods.

7) To evaluate the proposed plagiarism detection techniques by testing against an existing public domain corpus dataset.

This research is currently targeted tracing the identities of authors for many applications, especially in academic bodies and publication firms.

1.4 Research Methodology

This thesis has adopted a data driven methodology based on a combination of machine learning and stylometry techniques in order to address the problem of plagiarism detection. The machine learning algorithms require a dataset (described below) to perform the learning process (i.e. learning rules from data examples) in order to explore the correlation between tokens of text. The proposed approach employs statistical analysis of language features and uses a representative of each major class of machine learning method to validate the results.

For evaluation and validation of the proposed approach, the leave-one-out-cross-validation (LOOCV) method was used. This type of validation consumes a large number of computations but it is very accurate as the error rate is based on a single instance.

The application of this methodology involves several systematic steps; these steps are shown in Figure 1.1. The figure shows the main components for both types of approaches.

Data set

The dataset used for the research was the Corpus of English Novels (CEN). CEN composed by Hendrik De Smet whichhas been used in many studies as an example of sample texts from different authors. It is formed from English novels, written by twenty-five British (including Irish) and North American novelists. The novels were written in the period between 1881 and 1922,

furthermore all authors were born between 1848 and 1963 and represent roughly one generation of writers. The corpus is divided into training and testing data. The Bag of words method was applied to split the sentences into words which was then used for classification and model generation. In extrinsic plagiarism the suspicious text is compared to reference text which can help the proposed technique to identify the rules to discriminate between two pieces of text, All authors within the dataset represent the same generation which means convergent age, culture, environment and influential life factors.

Feature generation

Represents the first step of the procedure which starts by applying different techniques to prepare the including tokenisation of the raw text into words. It is also assumed as a pre-processing step to present a particular set of features, based on the task which needs to be conducted.

Feature selection

Builds a measured set of relevant features from the input data, eliminating those that are redundant. *Feature extraction*

After the initial step of features generation, feature extraction builds a set of informative features and eliminates the noise. A common example of feature extraction is the application of dimensionality reduction techniques which result in a reduced set of features to perform the proposed task. For the purpose of this approach specific pattern capturing processes need to be applied. The two separate legs of the diagram show the application of a slightly different set of procedures depending on whether intrinsic or extrinsic methods are required.

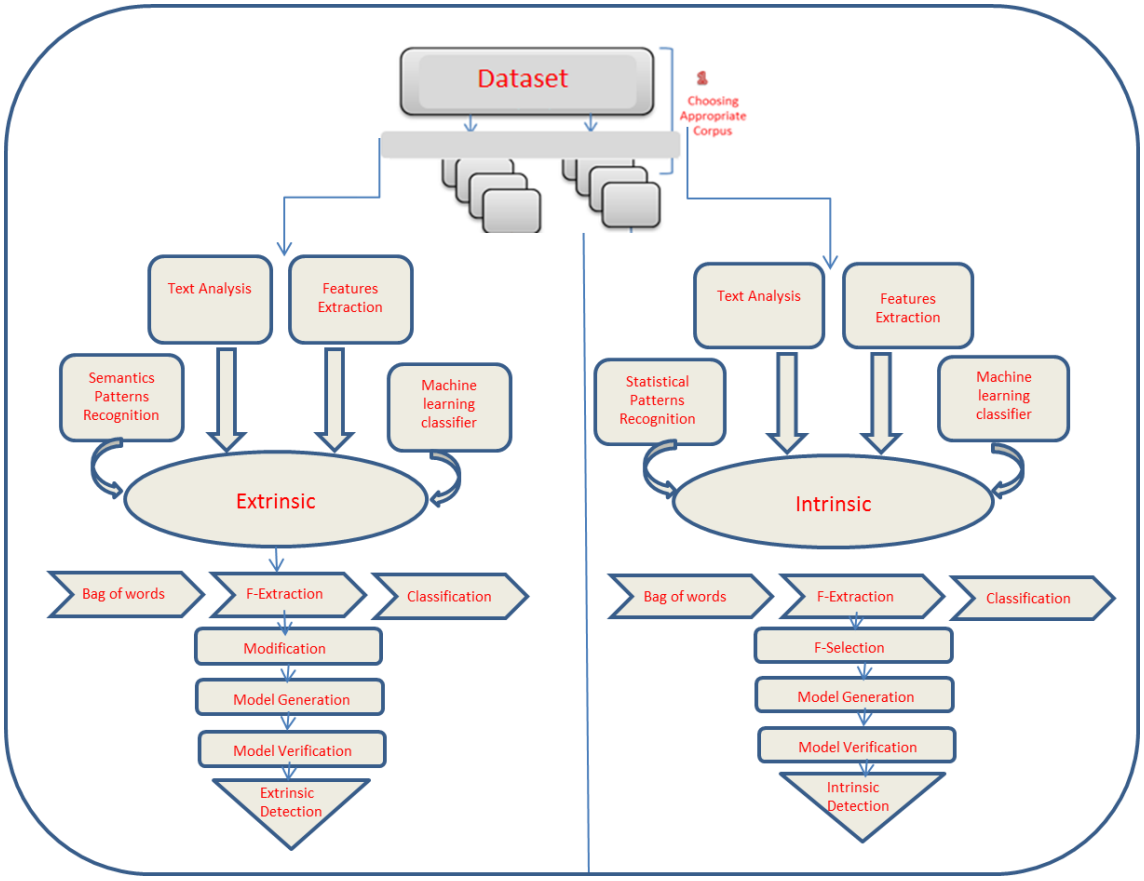

Figure1.1: The research methodology

1.5 Research Scope

The experimental design has used the statistical computing platform known as R, which is freely available for researchers. This research relies on the Corpus of English Novels (CEN) for experimental and evaluation purposes. Due to the scarcity of this research trend, leave-one-out-cross-validation has been employed on the CEN corpus. The findings of this research are limited to: the proposed data set; the experimental methodology; the parameters; and the techniques, which are used under specific conditions. However, the proposed approaches in this thesis are generalizable and can be applied to other datasets.

1.6 **Research Contribution**

This thesis has made a number of novel contributions as described in chapter 6 in detail and briefly outlined below.

1. The thesis has introduced a novel approach based on the integration of a number of well-known techniques in order to address the issues relating to plagiarism detection. More precisely the issue of text identification with and without a reference collection is identified and addressed by proposing two different but complementary approaches.

2. The identification and evaluation of the most suitable existing text analysis techniques to be used for by extrinsic and intrinsic methods for plagiarism detection. These techniques are discussed in Chapter 3.

3. A novel approach applying an extrinsic plagiarism detection method to reveal the semantic similarity between two texts and predict the authorship class from different classes by comparison with a reference collection. The approach consists of:

> (a) A new LSA application that was fine-tuned to take-in the stylometric features (most common words) in order to determine the document authorship. This is described in chapter 4.
>
> (b) A comparable class prediction machine learning technique. The technique was selected out of three different methods based on prediction accuracy. This is also clarified in Chapter 4.

4. A novel approach applying an intrinsic plagiarism detection method to verify the author of the target document where a reference collection is not available for comparison. The approach consists of:

> (a) A feature engineering method based on LSA and stylometry to produce a shrunken stylometric feature set which was used as a baseline.

(b) Sets of informative features were derived based on the MCW frequencies. The “in-series frequency ratios” of most common words outperformed all the other features. This represents one of the most original contributions of this work, as the “relative frequencies” of words (as opposed to the raw frequencies) have not yet been reported in the literature. This is clarified in chapter 6.

5. The research proposed a novel experimental methodology for testing the performance models using both extrinsic and intrinsic methods for plagiarism detection. The experiments covered a number of machine learning processes based on leave-one-out-cross-validation (LOOCV).

1.7 Structure of the Thesis

Chapter 2 discusses the relevant literature for the issues introduced throughout this thesis. The chapter reviews plagiarism detection challenges and existing detection models and reviews the techniques that were used in those models. It also presents a comprehensive review of both extrinsic and intrinsic plagiarism detection methods. The chapter compares the characteristics of extrinsic and intrinsic methods. The semantic analysis and text matching features of the extrinsic methods are reviewed. For intrinsic methods lexical and syntactic features are reviewed. A summary concludes the review and highlights the important key points.

Chapter 3 presents the background of the methods highlighted in the literature and discusses their empirical foundation. Four methods are discussed in this chapter, in particular those dealing with supervised learning. The key ideas for these methods are explored including: bag of words (BOW), latent semantic analysis (LSA), stylometry and machine learning (in particular those aspects of machine learning which deal with supervised learning). The underlying concepts of the method are discussed showing where they complement each other. For instance BOW is used to generate the initial features set by represent each word by a vector. As a result a high dimensional vector space is produced. LSA is used as a method for dimensionality reduction. On the other hand LSA

analyses the text content to reveal the text meaning, however it ignores the stylistic analysis.. In contrast stylometry is used to perform superficial analysis to capture text authorial attributes. Then machine learning algorihm is used for classification purposes.

Chapter 4 proposes a method for extrinsic plagiarism which addresses the gap in the research on plagiarism detection as briefly described in the literature review chapter. As highlighted by the literature, most automated plagiarism detection relies on string matching between the text and a reference collection. The semantic relationship between the plagiarised text and the reference document is ignored. Hence the semantics of the text is still a challenge for all supported plagiarism detection tools. The approach is based on cooperation between several techniques to accomplish the proposed approach task. BOW, LSA, stylometry and support vector machines (SVM) techniques are explored in chapter 3. The chapter discusses the components and the implementation details for each technique.

Chapter 5 presents a new intrinsic method for plagiarism detection. The key component of intrinsic plagiarism detection approaches is the ability to model the capacity of humans to detect variations in writing style. The method used the one-class classification method to conclude that test examples can either be classified to the target author (trained with correct labels) or a new class that was not available during training. In this scenario, two different predictions are possible: target, refers to examples that their class is learned during training, and not-target, where the xample does not classify to to the previously learned class. Four sets of statistical features are derived based on the frequencies of the MCW and their informative derivatives. Also a number of different machine learning models were generated, and trained using the features described above to evaluate the proposed approach.

Chapter 6 presents the results of the application of the two proposed extrinsic and intrinsic methods for plagiarism detection. The evaluation of the efficiency of these methods relies on a series of experiments on the corpus of English novels (CEN). The performance metrics of the extrinsic method was based on

the application of three support vector machines SVM based techniques. To evaluate the intrinsic method a number of different machine learning models were generated, and trained using the features described in chapter 5. This chapter highlights the most informative feature set which represents one of the most original contributions of this work, as this feature has not yet been reported in the literature.

Chapter 7 summarises the thesis outcomes and discusses the limitations and proposes directions for future research.

Chapter 2: Literature Review

2.1 Introduction

The previous chapter has discussed the motivation for this work by presenting the problem statement. The aim and objectives are also discussed in order to address the problem related to plagiarism detection. This chapter will further discuss the problem of plagiarism detection and present a number of existing approaches which aimed to address this problem with varying degrees of success. This chapter will also discuss the strengths and limitations of the existing approaches.

Plagiarism detection methodologies were stimulated by the authorship analysis approaches which use several text analysis techniques to infer the authorship of suspicious texts. In traditional authorship analysis, a suspicious text is attributed to one author, while given group of authors with their textual samples (Sebastian 2002). The authorship analysis approaches have stemmed from a linguistic root called stylometry which refers to the field of study analysing the author's writing style based on statistics by using computing algorithms (Abbasi, and Chen, 2008). Stylometry builds on a notion that each author has irreplaceable writing habits that cannot be imitated which are known as linguistic features or attributes (Burrows, 2002).

The current literature is embodied with a range of plagiarism detection techniques, although most of the available techniques have been broadly categorised into two: extrinsic and intrinsic (Stamatatos, 2009; Stein, Lipka, and Prettenhofer, 2011; Alzahrani, Salim and Abraham, 2012). Extrinsic techniques are most similar to traditional text classification algorithms, while intrinsic techniques use no direct comparison to an external document collection and are trained to recognise characteristic elements of the writer (Stein, Lipka, and Prettenhofer, 2011).

The literature refers to a wide range of plagiarism conduct types, starting from a simple copy and paste of the exact piece of text to the imitation of ideas, summaries or obfuscation by using translation (Afroz, 2013). Fig. 2.1 shows two types of plagiarism detection challenges: extrinsic and intrinsic. Extrinsic plagiarism detection challenges include text manipulation practices such as using synonyms of words, transferring from active to passive or vice versa. However, the challenges that faced intrinsic plagiarism detection were divided into two sub-practices: imitation and obfuscation. Figure 2.1 depicts the main unacceptable practices that may be conducted to confuse both types of plagiarism detection methods.

Figure 2.1. A classification of the plagiarism types and practices, the figure is stimulated by (Afroz, 2012)

This chapter explores the plagiarism detection techniques from a wide range of sources, drawing on both theoretical and empirical evidence. This chapter reviews the extrinsic and intrinsic methods for plagiarism detection, and explores the most informative features for capturing text semantics and text authorship.

This chapter is organised as follows. Section 2.2 explores the brief history of plagiarism detection methods. Section 2.3 describes the characteristics of extrinsic vs. intrinsic methods for plagiarism detection . Section 2.4 presents the summary of this chapter and highlights the main points that were surveyed.

2.2 A Brief History of Plagiarism Detection Approaches

Plagiarism detection systems actually began as detection tools for multiple-choice assessments (Angoff, 1974) and computer source code (Ottenstein, 1976). Prior to plagiarism detection in natural languages, code clones and software misuse detection had existed since the 1970s. At that time, a number of studies attempted the detection of plagiarised programming codes and algorithms (Ottenstein, 1976. Thereafter, plagiarism detection in natural languages through statistical or computerised methods began to gain popularity around 1990, founded by studies of copy detection mechanisms in digital documents. Between 1990 and 2000, most plagiarism systems developed were aimed at detecting programming code plagiarism, and only a few studies focused on plagiarism detection for written texts (Lathrop and Foss, 2000). A prototype known as COPS was an example of these early developed detection approaches for written texts, designed to detect partial or complete copies of digital documents using sentence-level matching (Brin et al., 1995). Although the strings of sentences in each document were matched against other sequences in the wider documents, this sentence-level matching approach seemed ineffective at detecting partial sentence overlaps. As a result of this inherent limitation, Shivakumar and Garcia-Molina (1995, 1996) proposed another prototype known as SCAM, as an extension to COPS. SCAM introduced, as a pre-processing step, the removal of both frequent words and stop words, instead ensuring the comparison of texts as overlapping sequences

of words or paragraphs. By doing so, thresholds are set which allows investigators to determine three levels of text overlap: minor overlap, major overlap and exact copies. This approach (i.e. using sequences of words as a feature) was found to outperform the previous ones and thus led to better accuracy, bringing about the notion of removing stop words being suggested as a direction for future investigation.

Another plagiarism detection tool that emerged at the time was the YAP3 tool (Wise, 1996), which was specifically designed to identify similarities in programming code. As a structured-metric similarity detection system, YAP3 utilised the Running-Karp-Rabin Greedy-String Tiling (RKR-GST) algorithm which is a modification of the Longest Common Subsequence (LCS) algorithm. This algorithm was designed to deal with cases where plagiarists have attempted the reordering of text sequences, which is possible because the tool allows a minimal match alongside a maximal match length between texts. Unfortunately, the YAP3 tool and the RKR-GST algorithm were mainly tested on computer source code. In other words, their effectiveness in written texts was yet to be verified at the time and further experiments were needed to evaluate this.

The recent developments in related fields such as machine learning, data mining, computational linguistics, and information retrieval (IR) has impacted the research on automated plagiarism detection in written texts. The impacts and development of the aforementioned approaches on two forms of plagiarism detection are discussed in the following sections.

2.3 Extrinsic vs Intrinsic Methods for Plagiarism Detection

Effective plagiarism detection is seen to be an essential capability in the next generation web. Plagiarism is now being applied intelligently, so imitation of the language should not be a target for the current detection methods. Ideas, methodologies and findings are being hacked and reproduced as new work without proper credit being given to the original author. Many commercial software systems (e.g., Turnitin©) were developed to tackle plagiarism. The

existing detection tools (commercially or freely available) are considered as text-matching software. They work on identical text comparisons and are all subject to comparisons with close world references (Stappenbelt and Rowles, 2010; Ramnial, Panchoo and Pudaruth, 2016; Eisa, Salim and Alzahrani, 2015).

The two most widely recognised techniques of plagiarism detection are: extrinsic and intrinsic (Potthast et al., 2009). With extrinsic mehthods, plagiarism is measured by comparing the suspicious document or string of text to a body of known classified documents which can be called reference documents (Alzahrani, Salim and Abraham, 2012). With Intrinsic methods, plagiarism is measured without reference to a set of known documents, using methods to infer the style of the suspicious piece, and judging from the results, deciding whether the style has been changed significantly (Potthast et al., 2009).

Yet, the majority of the existing detection tools (commercially or freely available) were categorised as extrinsic and criticised for their identical text-matching strategies (Stappenbelt and Rowles, 2010; Ramnial, Panchoo and Pudaruth, 2016). Turnitin© and CrossCheck, the leading plagiarism detection software in most academic institutions and publishing firms are still facing big challenges in detecting linguistic changes such as replacing words with their synonyms (Stappenbelt and Rowles, 2010; Eisa, Salim and Alzahrani, 2015). They are also criticised due to their vulnerability in increasing the numbers of false positives (the cases are detected as plagiarised but in fact they are not). As a result, they are always in need of human intervention to finalise decisions (Ramnial, Panchoo and Pudaruth, 2016).

2.4 **Extrinsic Methods for Plagiarism Detection**

As stated above, the extrinsic method for plagiarism detection relies on comparing the suspicious document or string of text to a body of known classified documents (Eiselt, and Rosso, 2009). Figure 2.2 presents the case of extrinsic plagiarism detection by determining the task and its requirements. Existing methods are inherently limited by their rigid assumptions that plagiarism is a copy and paste procedure (zu Eissen and Stein, 2006). This

assumption may be true for the lowest and most undefined form of plagiarism that is called the “first timers”, but it is certainly not true for other acts.

Figure 2.2. An example of an extrinsic method for plagiarism detection (Stein, and zu Eissen, 2007)

From the year 2000, the field of plagiarism detection has seen an increase in the number of new plagiarism detection tools, methodologies and methods of implementation. A significant number of scholars and research institutions began to pay more attention to the issue of written text plagiarism detection. This is evident from the growth in the number of commercial plagiarism detection systems available online, from as little as five in 2000 to about 47 in 2010 (Kohler and Weber-Wulff, 2010). Most of the existing plagiarism detection research mainly utilised non-NLP (Natural Language Processing) based approaches and detection approaches were largely insufficient in delivering the final outcomes on plagiarism cases (Eisa, Salim and Alzahrani, 2015). Human

judgement was often required in the end (Lukashenko, Graudina, and Grundspenkis , 2007; Ramnial, Panchoo and Pudaruth, 2016).

In a study that aimed to review existing plagiarism tools and technologies, Clough (2000) highlighted several related fields that may enhance the understanding of plagiarism detection. Clough (2003) also examined the nature of plagiarism in relation to the issue of multilingual plagiarism detection and suggested the use of machine learning methods and Natural Language Processing techniques as future improvements in plagiarism detection tasks.

Bull et al. (2001) evaluated five of the early plagiarism detection systems based on a technical assessment of their performances, using recommendations of the Joint Information Systems Committee (JISC). The evaluated systems include CopyCatch, Turnitin©, Findsame, WordCHECK and EVE2. The authors recommended carrying out further trials on three of the systems, namely: CopyCatch, EVE2 and Turnitin© in terms of improving their ability to handle larger datasets and to make detections from multiple sources.

In another pilot study to assess the use of Turnitin© in the educational setting, Chester (2001) recommended and approved of Turnitin© as an appropriate plagiarism detection tool for higher education institutions across the UK. However, subsequent general user feedback on the tool seemed unsatisfactory as Turnitin© lacked the ability to handle rewording or obfuscating texts effectively (Marsh, 2004; Weber-Wulff, 2008; Williams, 2002; Ramnial, Panchoo and Pudaruth, 2016).

Maurer et al. (2006) and Maurer and Zaka (2007) provided a comprehensive report on some of the challenges of plagiarism detection systems such as Turnitin© and Copycatch, and noted down how paraphrasing often renders these tools less effective. The authors revealed that existing commercial detection tools are largely unable to cope with synonyms, extensive paraphrasing and cross-lingual plagiarism, resulting in a number of plagiarism cases going undetected. They further recommended the use of an efficient algorithm to extract informative features before running a hybrid algorithm that

can work efficiently on the reduced dataset. This tolerates the reasonable application of integration between deep text analysis techniques and others which can be described superficially. For this research the extrinsic approaches are reviewed based on two categories: the text matching and semantic approaches.

2.4.1 Text Matching Approaches

Text matching or string matching approaches seek to detect the longest identical string between two texts. The detected strings indicate an attempt at plagiarism if the overlapped string exceeds the threshold (Gipp, and Meuschke, 2011).

Culwin and Lancaster (2001) introduced a prototype detection system that is capable of capturing the segments of plagiarised texts between two documents. The system was not able to detect plagiarism based on multiple sources. Advancing the initial research, Lancaster and Culwin (2004) subsequently investigated several matching methods and suggested that n-gram matching was the most efficient. The authors also developed and discussed their plagiarism detection tool known as PRAISE, which uses n-gram matching.

The n-gram overlap method is one of the most effective plagiarism detection approaches as it is based on calculating the number of word sequences between texts. This method has been applied across other fields such as text classification using 2-grams of words (Tesar et al., 2006). Text similarities are determined by distance or similarity metrics, which includes for example, the Jaccard index, Dice coefficient, Euclidean and Cosine distance. These powerful metrics generate similarity scores and rank documents based on their level of resemblance. For instance, with the aid of a simple string-matching algorithm, Monostori Zaslavsky, and Schmidt (2000) developed the Match Detect Reveal system to aid with the identification of direct copies of written texts. Although n-gram overlap methods have been proven to be effective in identifying direct copies, they tend to be incapacitated in circumstances where plagiarised texts involve more complex obfuscation such as paraphrasing.

A further example of the n-gram overlap method is the use of "overlapping 3-grams" as in the Ferret plagiarism detector (Lyon et al., 2001, 2006; Lane et al., 2006). The approach allows documents to be pre-processed into sets of 3-grams of words, comparing each set between document pairs. In this case, the similarity score is often determined by the Jaccard coefficient in which the number of matching 3-grams (which is the optimal n-gram size for matching shorter documents with minimal paraphrasing) is divided by the number of distinct 3-grams.

White and Joy (2004) discussed the use of text pre-processing techniques such as lowercasing, stop word removal, tokenisation and punctuation, and suggested the need to compare documents at sentence level. The sentence-based algorithm developed by these authors was able to detect direct copies, paraphrasing and sentence-level changes. However, the algorithm is only able to calculate the number of common words as well as the average length between the sentences; changes crossing the sentence boundaries are almost impossible to detect.

More recently, merging information retrieval models with similarity metrics has become popular in the field of plagiarism detection. For example, Tsatsaronis et al. (2010) investigated the use of similarity metrics with the Vector Space Model (VSM) and showed that, although using statistical metrics in detection tasks allows for simpler implementation as well as being effective against verbatim plagiarism, they often fail to aid semantic analysis of textual and non-textual information.

2.4.2 Semantic Detection Approaches

Alzahrani, Salim and Abraham (2012) have emphasised the urgent need for sufficient algorithms that are able to capture semantic patterns between two texts. A study by Ceska (2009) determined that existing detection tools are miscarried in detecting obfuscated text due to their algorithms' limitations. The similarity techniques of these algorithms do not take the linguistic or semantic structure into consideration during comparison procedures. The research in this

thesis has benefited from the work that was conducted by Ceska (2009) and followed the recommendation of integrating linguistic statistics and semantics to develop a new approach for extrinsic plagiarism detection.

Yerra and Ng (2005) proposed a technique that considers the text modification partially by replacing the text words with synonyms; the technique works on a specific set of alternatives. Another study by Alzahrani and Salim (2010) proposed a technique based on sentences to compare between two sentences using their words and corresponding synonyms. This method attempted to detect semantics; however, it needs to be improved to capture the correct meaning as not all synonyms relate to every meaning.

An investigative study by Sousa-Silva, Grant, and Maia (2010) examined the characteristics of plagiarism by analysing five Portuguese documents using a forensic linguistic approach. It was shown that replacing particular words with semantically related words, the insertion of words and a change of word order were major features that could confuse a plagiarism detection system.

A team of researchers applied the n-gram matching techniques for the favour of semantic detection. They assumed that two texts are semantically related when they share the same frequencies of text words, “semantic sequences”. The authors clarified that the semantic sequences procedure is a series of content words and no common or rare words were considered. The method has scored high accuracy in detecting exact copy and paste procedures but failed in detecting text obfuscation such as rewording or paraphrasing.

An approach to extract semantics was proposed by McCarthy et al. (2006) and was based on WordNet (Fellbaum, 1998). The approach was based on determining causative verbs (verbs refer to an event that will happen such as allow, let, help, etc.) and estimate information about synonyms and hypernyms (words’ categories; e.g. colour is the hypernym of red). They applied latent semantic analysis (LSA) to enhance semantics detection, however no details have been revealed about the features that were used or the approach outcomes.

Uzuner et al. (2005) targeted semantics through the use of semantic role labelling and rules to detect the re-writing of text pieces. They used the part of speech (POS) tagger to determine the verb's semantic category and analysed each sentence to identify the syntactic structure. They suggested that the string matching similarity measure was computed based on verbs' classes but not words. They applied their proposed approach to a translated dataset that includes 49 books to represent different types of medicated texts. Their results revealed that syntactic structure elements outperformed others and concluded that linguistic techniques can uncover paraphrasing. Uzner and his colleagues (2005) used a corpus that used naturally modified text to uncover re-written text which was assumed to be an advantage. However, Chong (2013) criticised the use of the corpus as she claimed that it retained the original text structure. Chong claimed that, when dealing with sentences, the plagiarisers always apply substantial structural changes.

Far from the above semantics detection attempts, there are just a few approaches based on Latent Semantic Analysis (LSA). LSA is one of the most well-known methods for semantics detection. It uses a mathematical internal algorithm to shrink the high dimensional vectors space. In addition, it works by revealing the latent association between words based on their co-occurrence (Deerwester et al., 1990). LSA has the ability to derive the connections between words by capturing the patterns of word usage (Landauer and Dumais, 1997). Researchers determined that LSA can extract the meaning from the text based on statistical computations as human do without complexity (Landauer and Dumais, 1997; Landauer, et al., 1998; Landauer, Laham, and Foltz, 1998). LSA is an intelligent method that is based on mathematical algorithms for text analysis. The method has a proven ability in revealing the underlying semantics in texts (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Dumais, 1991). Few researchers have applied LSA to plagiarism detection with a varying degree of success (Rehurek, 2007; Ceska, 2009; Britt, et al., 2004) for textual plagiarism and (Cosma, 2008) for source code plagiarism.

Rehurek (2007) suggested the use of a semantic-based approach by combining latent semantic indexing (LSI) and TF-IDF for the purpose of information retrieval purposes. He used the bag-of-words (BOW) technique and argued that BOW is not controlled by the sentence boundaries. In addition, this method maintains most of the relevant information. The researcher considers LSI as a derivative technique based on vector space models (VSM).

VSM focuses on analysing the conceptual relatedness between texts and explores the structure of co-occurrence words. Although this approach is promising, its results are not empirically supported so the evaluation standard is unknown.

Ceska (2007, 2009) proposed a method named SVDPlag which is based on Singular Value decomposition (SVD), the core of Latent Semantic Analysis. SVD is a mathematical internal algorithm works on decomposing the main matrix into three matrices to reveal latent associations. The method works on extracting features by reduction n-grams from textual documents; n was empirically evaluated from 1 to 8. The semantics correlations were embodied into the LSA model by a thesaurus in order to reserve the semantic relations between n-grams in text documents. Ceska also employed some pre-processing procedures such as lemmatisation and the removal of a set of common words.

The approach is considered promising in the plagiarism detection field, especially due to the use of LSA which was assumed as an untapped method for plagiarism detection tasks (Chong, 2013; Alzahrani, Salim, and Palade, 2015). However, the approach has relied on using SVD to reduce dimensions and the comparison was restricted to n-grams elements. The approach was described as a simple heuristic method by (Ching, 2013). This work has stimulated further investigation into plagiarism detection using LSA as well as other parameters that drive LSA performance.

Britt et al. (2004) have applied LSA to plagiarism detection in students' papers using their system which is called SAIF. SAIF has assumed an assistance system for students' development as it can give feedback regarding the lack of

citations and sentence modifications. The distinguishing feature for this system is integrating LSA with a group of techniques to accomplish the detection process. SAIF was able to detect around 80% of text misuse such as plagiarised sentences and incorrect citations. However, a comparative study conducted by Kakkonen and Mozgovoy (2010) to evaluate the state of the arts in the plagiarism detection area, demonstrated that the current detection methods are incapable of dealing with plagiarism challenges. They emphasised the most significant challenge in the plagiarism detection field is to deal with authorship detection. They also recommended that incorporating stylistic variation detection techniques with the current plagiarism detection approaches can substantially enhance the plagiarism detection performance.

WordNet and similar “online thesauruses” go a long way towards reducing the problem of synonyms. Several classes of tools such as information retrieval (IR) which is based on TF-IDF, multi-words, and in particular LSI/LSA (latent semantic indexing/analysis) exist whose goal is to identify the correct meaning of a word from the surrounding words, without using syntax and rising to the level and complexity of a semantic interpreter. A study from 2011 (Zhang, Yoshida, and Tang 2011) compared the three methods listed above (Wordnet, IR, LSA) using both an English and a Chinese test dataset. They found that LSI performed significantly better than the other two methods, and added that LSI has better semantic and statistical qualities. The superior performance of LSI may be due to a dimensionality reduction step for the word frequency vector, which removes a great deal of the irrelevant words. Research on text analysis for more than fifty years demonstrated that machines can understand the meaning of a word by identifying its co-occurrence with other words (Firth, 1957). This definition has opened up new research opportunities in the field of uncovering semantics. Sinclair (1991) and Louw (1993) discovered that words frequently occurring with other words can reveal important semantic patterns.

Some methods attempted to leverage statistics by using multi-word indexing terms (word n-grams) in place of individual words. Lahiri and Mihalcea (2013), for example, developed 8 distinct patterns of multi-word constructs and used

them in all 140 combinations, retaining 4 with significant information gain as the final features for a predictive model of language identity. Although this method is reasonably successful, it has not demonstrated any clear advantage over the more established latent semantic indexing (LSI), which uses single-word indexing. In fact, no study has convincingly shown multi-word indexing by itself to offer an advantage over single-word indexing (commonly known as “bag of words”).

The fundamental problem that literal, word-for-word, phrase-by-phrase comparisons are being used to identify plagiarised text remains which is assumed as insufficient. In short, it is relatively easy to evade detection by simply rewording or re-organising the text.

The literature has clarified that plagiarism detection is more complex than just identifying copied and pasted plagiarised pieces of text (de Jager, and Brown2010)

2.5 Intrinsic Methods for Plagiarism Detection

The previous section has described the existing studies in the field of extrinsic methods for plagiarism detection for text documents, concentrating on methods that compare suspicious text against references set or classified documents. While these methods perform well to some extent for copy and paste misconduct, the detection assumption was built on a notion that all related information is digitalised. A criticism has been raised against that assumption revolves around the fact that not all sources are digitalised (Eissen, Stein, and Kulig, 2007). Consequently, a new class of plagiarism detection tools is currently being researched and developed, termed “intrinsic” detection methods, because they aim to characterise a writer’s style using a history of that writer’s existing work (Zechner et al., 2009). The intrinsic methods for plagiarism detection differ from extrinsic methods as they do not use a references collection to compare with, as shown in Figure 2.3. This type of detection method relies on capturing the variations in written text by extracting the syntactic and lexical features. Then a comparison is performed

between the suspicious text and the same author's work in order to identify the variation patterns.

A wide range of plagiarism and authorship analysis approaches were developed as presented in the above survey, using a diverse set of features and text analysis techniques, however no standard platform was proposed in order to compare their performance. These approaches were examined in different datasets. Hence their performance cannot be compared and the results cannot be generalised.

For centuries, scholars have sought to find more reliable ways to prove the authorship of certain important documents. Even scholars who have spent a lifetime analysing certain documents and authors often did not agree on authorship (e.g. a number of works generally attributed to Shakespeare, but are argued to have been written by Marlow instead (Zhao, and Zobel, 2007). These tools can help to generate a model of the author's style and help to reveal certain features of authorship (e.g. for literary analysis). However, these tools are normally evaluated based on a small dataset (Luyckx and Daelemans, 2008).

The early basic set of techniques in authorship analysis has relied on selecting features from an author's written texts that are unique to that author (unitary) and these features do not change over time (invariant). These techniques were discussed and defined in the late 19th century by Mendenhall (1887) who studied the texts of Shakespeare, as well as Marlowe and other contemporaries.

Figure 2.3. Presents the task of intrinsic methods for plagiarism detection (Stein, and zu Eissen, 2007)

Mendenhall ultimately discovered that a characteristic can often be found by plotting the curve of frequency vs. word length for a particular author. These two characteristics have established a foundation for the characterisation of an author's writing style and formed a strong basis for statistical approaches (Stamatatos, 2009). Researchers have continued to search for a single feature that is unique for a specific author and unchangeable during the time. Many suggestions for such features were established such as calculating the average word length suggested by Fucks (1952). Another suggestion by Yule (1944) to calculate the average number of words in each sentence was proposed. These calculations and their counterparts are considered as insufficient to identify text authors (Koppel and Schler, 2004).

The next most sophisticated development in the history of stylistic methods which represents the next generation class of features originated with the most common words (MCW) which can sometimes be called function words

(Stamatatos, 2009). Function words are language elements without (much) inherent meaning whose primary purpose was to clarify the relationship between words' classes in different textual parts. An analysis of these parts and the statistics of the most common words has remained a popular topic ever since its inception in 1964 by Mosteller and Wallace. The comprehensive study by Mosteller and Wallace (1964) was to investigate the authorship of 146 political articles written by James Madison, Alexander Hamilton and John Jay. This was an important milestone work in the field of authorship analysis. This issue was named The Federalist Papers disputation as twelve of these articles were claimed to be written by Madison and Hamilton. The study found that measuring the frequencies of a specific set of the most common words could result in improving the prediction of the text author, compared to content words or other word classes. One logical explanation for the outcomes of this study is that the unconscious use of a set of words remains constant, even when the topic changes. The study of Mosteller and Wallace was assumed as the solid foundation for using statistics in authorship analysis. In addition, the application of the study has led to the birth of stylometry (Gruner and Naven, 2005). This thesis proposes approaches which were stimulated by the statistical analysis that was conducted by Mosteller and Wallace (1964).

Another study conducted by Farringdon (1996) to apply statistical analysis resulted in developing a system named CUSUM. CUSUM is a system that was used in the field of authorship analysis to form the accumulation sum of the abnormalities of the measured features. The results have been plotted to compare the performance of analysing the authorship and it was considered as a capable detection system. CUSUM turned out to be a forensic analysis method to provide forensic experts more evidence to decide about detecting the right authorship. Two years later Holmes (1998) criticised the performance of CUSUM as its result cannot be trusted over different topics; changing the topics affected its performance negatively.

Holmes (1998) confirmed that the use of the most common words is more effective in distinguishing between authors as each author has a unique usage

pattern of this class of words. This hypothesis was also confirmed by several authors such as Merriam and Matthews (1994), Juola and Baayen (2003), Koppel, Schler, and Zigdon, (2005), Zhao and Zobel (2005) and many others.

An important application using the most common words was adopted by Burrows (1987) who applied principle component analysis (PCA) on a set of the most common word frequencies. PCA was able to connect a wide range of measures and project them into a graph to measure the similarity distance between several authors. This trend in research encouraged other researchers to follow Burrow's procedure. Biber (1995) applied a statistical method to describe variability among features. The method was called factor analysis. Biber used this method to discriminate between four texts' languages.

The advent of machine learning algorithms in the authorship analysis research field influenced the research movement. Multilayer perceptron or the artificial neural network algorithm was employed by Tweedie et al. (1996) for the authorship analysis task. Tweedie and his colleagues used three hidden layers to train the political articles with a conjugate gradient and two output layers. They reported that the results were harmonised with the previous studies which worked on the same articles. Support Vector Machines (SVM) was introduced by Diederich et al. (2000) to recognise the stylistic features of seven authors. The dataset includes 2,652 newspaper trainings written by several authors covering three subjects, with a detection accuracy ranging from 60% to 80%.

Depending on the previous studies, textual features taxonomy has been developed for authorship analysis tasks by Zheng et al. (2006). Figure 2.4 presents four types of feature sets; each set includes a group of influential features that can affect the performance of detection approaches. The textual features types are: lexical, syntactic and structural and content specific. Structural features include paragraph length, use of signature and specified indentation. These features were considered as discriminative authorial attributions for authors' writing style. Such features strongly depend on the person's writing habits. Lexical features include the frequencies of any class of

words based on the predefined task and also the punctuation frequencies as shown in Figure 2.4.

The syntactic features can be defined by the function word usage, punctuation usage and part of speech usage, POS. Finally, the content specific features are the words that related to a specific domain and keyword frequencies. The following Figure 2.4 presents the feature types and gives examples for each type.

Figure 2.4. Captures the features types' taxonomy and the most important related features, the figure was stimulated by the study of (Zheng et al., 2006)

Koppel and Schler (2004) proposed a one-class classification method in order to identify if a specific written text was written by a target author or not. The method works by stemming two pieces of text and then analysing them using computational stylistics. Based on the analysis, it decides if these two texts

were written by one author or more. They concluded that the use of negative examples in the language model influenced the classification accuracy.

In the same context, Koppel et al. (2009) identified three types of scenario for their approach to be performed. They proposed a classification procedure to detect the author of a text when there is no candidate corpus so they analysed the writing based on age, education level, and gender and so on. In the second scenario, they assumed there is a large number of authors (thousands) and the available sample of text for each is very scarce. Koppel's group described this scenario as searching for a "needle-in-a-haystack". In the third one, which the assumption is based on, there is no closed references set to compare with but there is one suspicious set. This is called authorship verification or intrinsic plagiarism and the challenge is to decide if the suspicious text is the author or is not. They concluded that Bayes and the Support Vector Machine (SVM) performed better in the context of their experiment.

Meyer zu Eissen and Stein (2006) applied the stylometric method by calculating the average number of sentence lengths for all documents and calculating the number of word classes such as nouns, adjectives). They also calculated the frequencies of special words (frequent and rare words) and their average frequencies. All features were extracted from both the suspicious and original documents to form the input variable sets for machine learning. Meyer zu Eissen and Stein (2006) also stated that the feature sets were analysed using SVM for classification tasks. They reported that the average word frequencies and the average sentence length outperformed other feature sets. A subsequent study by (Zu Eissen, Stein, and Kulig, 2007) used the same feature sets from their previous study to investigate the performance of their approach by measuring the vocabulary richness. They calculated the vocabulary richness by dividing the average word length by the sentence length. They used a dataset of fifty documents written in German that artificially partially plagiarised and then employed them in a linear classification algorithm. They followed the method that was applied by (Zheng et al., 2006) by applying a feature combination instead of an individual set.

A major drawback of intrinsic plagiarism detection methods and their parent class of stylometric methods is that they rely on such a small training data set. Typically, the number of documents available for an author under suspicion of plagiarism is fairly small. Style is dependent upon not just the author, but also the level of technicality expected from the work, the length of the work, purpose, and degree of formality. Use of the first person subject and imperative tense would affect a persuasive essay; however, this would not be the case for a scientific report. The small sample size plagues all stylometric methods, and likewise all intrinsic plagiarism detection methods.

In order to overcome the small sample size data set problem inherent to the stylometric method style of analyses, researchers have proposed various different kinds of text and document features as shown in Figure 2.4. These range from simple (tokens such as word length, word per sentence, and other distributions), to higher level (syntactic features such as frequency of the passive voice, nominalization count, and distributions of frequency of different parts of speech tags), to expert-based (measures of vocabulary and rare word richness). Zu Eissen and Stein (2006) proposed the use of more standard and word-frequency features to develop a plagiarism-detection method called a “taxonomic tree”. The tree presents the taxonomy of the misconduct as shown in Figure 2.5. This also shows the specific part of plagiarism detection without the available corpus to compare with.

Figure 2.5. Taxonomic tree of plagiarism-detection methods according to reference document collection size, style of text analysis, and stage in the plagiarism detection process (i.e. processing of accurate copy vs. modified copy) (zu Eissen and Stein, 2006)

As briefly stated above about stylometry, it has been defined as the linguistic root for assigning a text to its reliable author. This method is based on a statistical analysis of an author's writing style (Craig, 2004). The textual feature types include lexical, syntactic and structural and content specific, however each study can have different feature types based on its aim and experimental corpus.

2.5.1 Lexical Features Related Studies

Vectors of word frequencies are considered as the clearest and most straight forward method in representing the text; most authorship analysis studies relied on lexical features to represent the writing style. Stamatatos (2009) described a method to select lexical feature sets for authorship analysis by calculating the frequencies of the most common words in a corpus. After the initial calculation to explore their landscape, the number of MCW which need to be selected is always rely on practical basis. This representation relates to the bag of words (BOW) text representation technique in the majority of classification approaches

(Sebastiani, 2002). In the BOW technique, each document is broken down into individual words followed by their frequencies without giving attention to the word order (Wallach, 2006).

Burrows (1987) argued that a sub-set of words such as the most common words (MCW) was considered a significant discriminator between different writing styles of authors. However, such words are often omitted during the pre-processing stages in most classification procedures as they do not handle any meaning and they are always called function words (Stamatatos, 2009). On the other hand, in authorship analysis, classification methods based on lexical procedures require much lower dimensionality features compared to other classification procedures. Furthermore, many researchers have recommended employing function words for stylistic based text classification because they are assumed to be topic-free. These words are also used by authors in an unconscious way so they cannot be imitated (Boukhaled and Ganascia, 2015).

For authorship analysis based on the English language, different sets of function words and other classes of words have been utilised. However, insufficient details are found in the literature. A number of different classes of words of interest have been identified. One of the most successful approaches is also one of the simplest, looking at the words with the most information gain, in a mutual loop by which information gain is defined by the ability to correctly attribute authorship in a training corpus. Quinlan (1986) found this to be a very useful approach, identifying the 1,000 most informative words in the 10,000 most common words of a corpus of English novels. In addition to the selection of features to use in combination as a predictor of authorship, multivariate approaches also employ different means of comparing documents across these (often quite large) feature sets. Whereas with a unitary invariant (unchangeable over time) feature, the standard deviation across the author's documents can be used, but with multivariate features more complex statistics are needed. Argamon, Šarić, and Stein (2003) used a set of 303 words. Abbasi and Chen (2008) used a set of 150 function words. Zhao and Zobel (2005) used a set of 365 function words. Another set of 480 function words was suggested by

Koppel and Schler (2003) as well as another set of 675 words proposed by Argamon et al. (2007). Burrows (1987) and Lampson, Abadi, Burrows, and Wobber (1992) used the 100 most common words and described this number as accurate to characterise the text authorship.

Burrows (2002) proposed an approach which is based on Laplace distribution. The approach helps us to discover the maximal probabilistic ranking of a set of features which is known as the probabilities of features for particular authors. This approach led to the development of a number of other, related distance measures for predicting authorship based on multiple features. The method uses z-scores values by calculating the distribution of 150 MCW, then the standard deviation for each word in the document to explore its usage in the document if it was used more or less compared to an average. Burrows (2002) chose a range of words to form a foundation for stylometric comparisons.

Rather than determining the number of function words that need to be used in such an analysis, the classification algorithm also needs to be investigated to avoid the overfitting problem when the features' dimensions increase. The advent of machine learning algorithms in the field of authorship analysis reduced the impact of a large number of features' dimensions. Joachims (1998) determined in his research that SVM can deal with a high number of features without a high risk of overfitting. Madigan et al. (2005) have proposed a brute-force technique using all words that appeared twice or more in the dataset. They concluded that the first dozen MCWs are dominated as stylistic informative, and after hundreds the open-class words started to appear. As a result, the large number of dimensions far from the classification algorithm effectiveness affected the stylistic detection as content words started to appear.

As stated above, the BOW technique can offer an efficient text representation manner for the initial stage of text analysis. It was confirmed by Coyotl-Morales et al. (2006) and Sanderson and Guenther (2006) that classification accuracy was significantly enhanced when individual word features were employed. Over

the years, enormous efforts have been made to enhance or replace the BOW representation with useful features (Sebastian, 2002). Boulis and Ostendorf (2005) argued that the old methods that worked to improve BOW did not achieve satisfactory outcomes as they just considered the relevant features. The noise features also need to be considered and a careful strategy needs to be used to keep the relevant features and features that are jointly correlated.

The n-grams approach which is widely used in authorship analysis techniques can be identified as a series of adjoining words usually known as word collocation (Peng, Schuurmans, and Wang, 2004). This method is criticised by many researchers such as in Sanderson and Guenther (2006). They stated that the n-grams method is a useful contextual feature for authorship analysis methods. On the other hand, Gamon (2004) described that text representation using n-grams is very sparse due to their internal technique which allows all possible groupings between words. In fact, no study has convincingly shown multi-word indexing by itself to offer an advantage over single-word indexing (commonly known as BOW).

Remarkably, in real world authorship analysis, human experts use similar methods (e.g. quantifying some stylistic features) in order to characterise the anonymous text authorship. Nevertheless, the availability of a precise detection method is still challenging in a wide range of tasks in authorship analysis.

2.5.2 Syntactic Features Related Studies

Most surveys that have targeted stylometry have found that syntactic features are essential for authorship analysis in varying degrees. A study conducted by Holmes in 1998 pointed out that the use of syntactic features has increased due to the advent of the digital era in 1990. Another study by Stamatatos et al. (2000) highlighted the growth of using syntactic features. Stamatatos (2009) claimed that syntactic features are based on an idea that authors commonly use some words unintentionally. This attribution has credited syntactic features to be more distinctive than lexical ones (Hollingsworth, 2012).

The syntactic features that were used for the purpose of stylometric analysis in authorship applications are assumed to be superficial text analysis features. Many researchers have proposed their own methodologies to measure the variations between different authors' writing styles. Examples of such syntactic features can be summarised by the sentence length feature which was proposed by (Yule, 1944). Also, the distribution of POS is another syntactic feature and then a particular class of word ratios was used as well (Stein and Meyer zu Eissen, 2007). Function words or the most common words have attracted attention as useful syntactic features since they are used unconsciously to represent certain syntactic structures (Ganapathibhotla, and Liu, 2008; Burrows, 2002; Hollingsworth, 2012). The achievements of MCW in the field of authorship analysis as a salient stylistic feature highlighted the usefulness of syntactic information (Abbasi and Chen, 2005; Burrows, 2002).

On the other hand, chunking the texts into different types of tokens has given the chance to construct different types of attributions by applying the relative frequencies. This participates in developing more accurate natural language processing (NLP) techniques (Chaski, 2005). Other researchers have suggested that the use of frequencies of short text chunks and integrating them with other word classes can influence the outcomes of syntactic features for text authorship analysis (Koppel, Argamon, and Shimoni, 2002) and (Zheng et al., 2006).

Koppel and Schler (2003) have proposed an interesting approach to scan the syntactic errors such as incompatible tense. They applied commercial spell checker software to track the errors. This method is considered as an imitation technique for human experts' ability in order to carry out authorship analysis tasks. They concluded that the spell checker software was not precise in detecting such errors and they needed to adjust the software outcomes to improve the detection performance.

Some studies recommended a combination of features or strategies to deal with authorship applications. They claimed that the selective syntactic features are

considered less useful when applied separately to address the writing style characteristics (Chong, 2013) and (Hollingsworth, 2012).

2.6 **Summary**

The above discussion targeted the plagiarism detection methods in both forms: extrinsic and intrinsic. The research in these methods is expanding and developing at an explosive rate. This growth has been fuelled by demand from the growth of the World Wide Web, social networks and online services. Plagiarism is on the rise, and has become a major concern for many sectors. However, plagiarism detection can no longer be performed manually by humans; there is a need for efficient methods. Existing detection methods are considered insufficient in their performance because they are inherently limited by their rigid assumptions that plagiarism is a copy and paste procedure. The existing plagiarism detection tools (commercially or freely available) were described by many authors as text-matching software. They work by comparing the suspicious text with references collection to detect the identical similarity (Ramnial, Panchoo and Pudaruth, 2016). A comparative study conducted by Kakkonen and Mozgovoy (2010) to evaluate the state of the arts in the plagiarism detection area demonstrated that current detection methods are not able to deal with plagiarism challenges. They highlighted that the most significant challenge in the plagiarism detection field is to identify the text source. They recommended that incorporating authorship detection techniques with the current plagiarism detection approaches can substantially enhance the plagiarism detection performance.

The key area of research interest is now to extend authorship analysis methods and other text-classification techniques into plagiarism detection in novel manners that confer advantages over the previous methods. As the literature has split the plagiarism detection methods into two forms: extrinsic and intrinsic, the above survey has explored the method related to them. Extrinsic methods rely on comparing the suspicious text with references text collection using a wide range of text analysis techniques. Researchers have noted down that rewording the text often renders these tools less effective. They revealed that

the existing commercial detection tools are largely unable to cope with semantic plagiarism challenges. As discussed above, LSA can improve both the performance and runtime efficiency of text-classification, as was evidenced in its ability to reveal semantic patterns between two texts. Intrinsic plagiarism detection methods use no direct comparison to an external document collection. Instead, they use stylometric analysis to generate the characteristic features of an author's body of work, to be compared with suspicious text. The disadvantage of intrinsic plagiarism detection methods is the lack of a data set. Stylometry is assumed to have a linguistic root for all authorship analysis applications. It builds on a notion that each author has irreplaceable writing habits that cannot be imitated which are known as linguistic features or attributes (Bailey, and Ottaway, 1979). Over the years researchers have been keen to find a single feature that is unique for a specific author and unchangeable over time. Many suggestions for such features were established such as calculating the average word length (Fucks 1952) or the average number of words in each sentence (Yule, 1944). All these suggestions are considered to be insufficient to identify text authors (Koppel and Schler, 2004). The next most sophisticated development in the history of stylistic methods which represents the next generation class of features originated with MCW (function words sometimes) (Stamatatos, 2009). Function words are language elements without (much) inherent meaning whose primary purpose was to clarify the relationship between words' classes in different textual parts. Many studies, as explored above, have recommended the use of function words as they provide topic-free features and are used by authors in an unconscious way. Hence they cannot be imitated (Boukhaled and Ganascia, 2015).

The use of stylometry and machine learning techniques in plagiarism detection is still scarce. Classification techniques relied on using the BOW technique for text representation. Based on BOW, each document has been broken down into individual words followed by their frequencies without giving attention to the word order (Wallach, 2006). It was confirmed by Coyotl-Morales et al. (2006) that classification accuracy is significantly enhanced when individual word

features are employed. Sebastian (2002) determined that the majority of text classification techniques uses BOW.

It is evidenced in the literature review chapter that Latent Semantic Analysis (LSA), stylometry, machine learning and bag of words have good potential to address the plagiarism detection when integrated together. However, these methods have their own limitations which need to be addressed in order to produce better results. This research was inspired by a number of studies that were explored in the literature; the frequencies of the most common words are used as a mean parameter in the proposed approaches. Burrows (1987) argued that a sub-set of words such as the most common words (MCW) was considered a significant discriminator between the different writing styles of the authors. However, such words are often omitted during many pre-processing stages in most applications. Ceska (2008) used LSA for semantic plagiarism detection which is assumed to be an untapped method in this field. The study of Zheng et al. (2005) stimulated the method of applying the machine learning methods and feature sets.

The next chapter will discuss these methods in detail in order to develop further understanding of their underlying concepts and to assess their feasibility to be used as part of the proposed approach.

Chapter 3: Background

3.1 Introduction

The previous chapter surveyed the plagiarism detection approaches from a wide range of sources, drawing on both theoretical and empirical evidence. Several text analysis techniques have been reviewed based on their contributions to both extrinsic and intrinsic methods for plagiarism detection. The detection process is performed by using text matching techniques between a pair of texts in order to highlight similarity. The highlighted cases then are taken as indicators for potential plagiarism action, requiring further human intervention or investigations (Lancaster and Culwin, 2001).

This chapter provides a review of the key ideas to the aforementioned methods that were highlighted by literature review chapter. It also sheds light on each method and discusses their characteristics in order to enhance the empirical integration foundation.

It is evidenced in the literature review chapter that Bag of words (BOW), Latent Semantic Analysis (LSA), stylometry and machine learning algorithms have a good potential to address the plagiarism detection limitations. However, these methods need to be integrated in order to effectively address the limitations of existing plagiarism detection methods.

This chapter is organised as follows; section 3.1 explores BOW as an initial text representation technique. Section 3.2 reviews how LSA works, and parameters that drive it were considered by this research. Section 3.3 discusses stylometry, defines its concept and describes the salient sets of related features. Finally section 3.4 discusses the machine learning methods in general and explores two of them; Support Vector Machines (SVM) and Multiayer Perceptron (MLP). The conclusion of this chapter is presented in section 3.5

3.2 **Bag of words (BOW)**

BOW is one of the popular text representation techniques that is used to represent text in many applications in particular text classification. BOW relies on a notion that each word establishes a dimension in a vector space isolated from any other words (Salton, Wong and Yang, 1975). Researchers have reported that the BOW method can perform much better when integrated with dimensionality reduction methods. This recommendation was proposed because BOW follows a strategy that each word has its own representation in vectors space with no connections to other words (Altınel, Ganiz and Diri, 2015). The functionality of BOWs is to break documents into unique words, counting the frequency of each term to form the "baseline" features'. Word counts are important because they form the basic input for a common class of text classification technique. BOW is based on the score of all the individual words stored as a high-dimensional vector of frequencies.

The main functionality of BOWs in the proposed research is to break documents available in the corpus into unique words, counting the frequency to form the "baseline" features' set as shown in table 3.1. The table describes how BOW works.

Table 3.1. BOW representation, the number of books and terms T1.Tn have been used as examples

	Author 1				
	Book 1	Book 2	Book 3	Book 4	Book 5
T1	2		4		1
T2	1		2	1	
T3		1		1	
T4		1	2		3
T5	3	2		2	1

In table 3.1, each row represents a unique word or term (T) and each column represents a document (book). Each cell entry represents the frequency of each word in each book, for example in the first row T1 appears 2 times in Book1 while it appears 4 times in Book3.

In the literature plagiarism detection methods were described to be used in a similar manner to the general text classification methods and BOW is evidenced by the literature as a complimentary method for text classification. This means that BOW can be considered as a robust and well performing starting point in terms of a first step feature generation for plagiarism detection approaches.

3.3 **Latent Semantic Analysis (LSA)**

Latent Semantic Analysis (LSA) is an intelligent technique that captures latent semantic associations based on the usage of words (Landauer, Foltz, and Laham, 1998). This method has been used as an information retrieval technique (Edmunds 1997), and lately for plagiarism detection (Cosma, 2008; Ceska, 2009). It works in deriving measures of the similarity of meaning between words from the text to mimic human word sorting and category judgments. LSA does not use any linguistic elements such as grammar, syntactic parser or dictionaries.. However it depends on parsing the raw text into words and works on extracting the semantic association based on pure

mathematical model called Singular Value Decomposition (SVD) ((Landauer, Foltz, and Laham, 1998).

LSA performs the following four steps:

1. A pre-processing includes tokenisation of the raw text into words. After that two more processes take place; stop-word removal and stemming of words. Stop-word removal procedure involves in eliminating words with a high frequency such as prepositions, conjunctions and common words from corpus' documents. While the stemming returns the words into their roots by removing suffix and prefix.

2. After the first step (pre-processing procedure), a matrix of m × n is constructed. Each row represents the actual frequency of a single word and each column represents a document where the frequency of a single word appears.

3. A preliminary transformation procedure is then applied by giving a suitable weight for each cell entry to express the word's importance in particular document (local weight) and in the whole corpus (global weight) using TF-IDF weighting formula. This formula gives low weight for MCW, prepositions and so on.

4. Once the weighting procedure is completed, the mathematical method; the Singular Value Decomposition (SVD) is performed to derive the model of semantic structure. The process of SVD in LSA works in the following manner and as shown in figure 3.1. Assuming A is the original matrix of document-word associations, of dimension

m x n:

$$A \approx A_k = T_k S_k D_k^T \qquad 3.1$$

Figure 3.1. Decomposition procedure using SVD, this figure was stimulated by (Deerwester et al., 1990)

Where, *T* is an *m* x *r* version of *A*, *S* is an *r* x *r* matrix of singular value, and *D* is an *r* x *r* matrix of document vectors. The *k* subscript refers to the uniquely LSA contribution to the SVD operation, *k* is the number of singular values taken to represent the original term-document matrix, usually fairly small for efficient computation (and effective noise-reduction). The parameter *k* is smaller than the rank of the matrix *A* (≤min (*m*,*n*)). Although the value of *k* should be kept fairly small for efficiency, the parameter is adjustable so that LSA allows for a choice to be made with regard to informational richness (or rather, the trade-off between informational richness and computational efficiency) (Berry, Dumais, and O'Brien, 1994; Berry, 1992).

The tables 3.2 and 3.3 clarifies the values of co-occurrence words before and after truncated the main entry matrix by SVD as shown in formula (3.2). Table 3.2 shows the co-occurrence values of terms; table 3.3 shows the approximation values after applying SVD to a specific value for k. Remarkably, as can be seen in table 3.3, user and human terms now have a value of .94, representing a strong correlation, where the earlier value was zero. In fact, user and human is an example of second order co-occurrence.

Table 3.2. Represents the relation between word & word before applying SVD as stimulated by (Deerwester et al., 1990)

Table 3.3. Represents the relation between use & human after applying SVD to a specific k value (Deerwester et al., 1990)

LSA is considered a highly parameterised statistical technique as the performance of LSA is driven by its parameters (Jorge-Botana, Ricardo and Escudero, 2010). The parameters' settings are varied from task to task as they need to be adjusted for a specific task to which LSA is applied (Cosma and Joy, 2012). The following section discusses the parameters that are related to this thesis which includes:

1- The value of k: This represents the number of singular values that applied to reduce the original post-SVD matrix. Identifying the ideal number of

dimensions to preserve is still a challenge as no two researchers used the same number of dimensions. As a result, k value assumed as an empirical number and still unanswered as researchers used different values of k's depending on context of their tasks (Jorge-Botana, G., Leon, J.A., Olmos, R. and Escudero).

2- Weighting methods refers to the procedure of transforming each word frequency in the matrix, using local and global weighting formulas. This parameter can play a significant role in enhancing LSA performance. It is argued that the procedure of weighting a group of words depends on a specific task and dataset which is used in the experiment (Berry, 1992).

As clarified above the application of LSA relies on a series of processes; pre-processing, transformation and constructing the semantic space. The MCWs are commonly eliminated during the pre-processing course (Cosma, 2008; Ceska, 2008). Even if they have been retained they will be ignored during the transformation step because they will be given trivial weight values. In many application areas, LSA is designed to ignore the MCW due to their high frequencies but with no contribution to the text meaning. On the other hand the MCW are assumed as a salient stylometric features which can play significant discrimination roles between classes. This contrast between LSA and Stylometry can be adapted in the proposed research for authorship analysis.

3.4 **Stylometry**

Stylometry relies on the use of computational algorithms to analyse the writing style of a specific author statistically. This method is used to uncover the variation between two pieces of texts and assumed each author has an inimitable writing practice that is conducted unconsciously. These inimitatable writing practices can be computed to create a unique writing style for each author in order to compare with others. These unique writing features are measured to create an author profile against which other texts or authors can be compared (Batineh, 2015). It is well-known method and is widely used in different applications such as forensic analysis and authorship studies to assign

a piece of text with evidences to a specific author based on stylometric quantification.

A study conducted by Oakes, and Ji, (2012) has claimed that stylometric analysis is considered to be one of the most trusted procedures in recent years. They also argued that stylometry can be used to analyse the authors writing styles and extracted informative features for best authorship analysis practices.

A unique word usage pattern can be captured for each author, to generate a signature recursive pattern for the text. One of the most stylometric salient features is to use the frequency of function words to quantify stylistic features (Mosteller and Wallace, 1964; Burrows 2002; Stamatatos, 2009). The use of function words is considered to be the best discriminant approach in Stylometric methods owing to its independence from topics and its unconscious use by authors (Stamatatos, 2009; Boukhaled and Ganascia, 2015). Currently most researchers depend on computational processes instead of linguistics and literal procedures (Stamatatos, 2009). There are two scientific disciplines categorised in computational processes to deal with textual features; artificial intelligence (AI) represented by artificial neural network (ANN) to deal with stylometric specific tasks. The computer aided statistic techniques such as LSA and PCA are applied for dimensionality reduction and patterns capturing for some applications (Rehurek, 2010; Ceska, 2008). Capturing the text patterns is based on the transitivity rule, these techniques help to capture specific words' occurrences in text and reveal latent associations. In this research two main stylistic features were proposed; content-words (CW) and most common words (MCW).

3.4.1 Content-words (CW)

CW are artefacts of particular writing attitudes or belong to specific topics, such as nouns, verbs, adjectives, and adverbs which describe some objects, actions, or statuses. They contrast with the MCW in functions, furthermore they are considered to characterise text context but not author's style (Koppel, Schler and Argamon, 2009). A well-known study by Maurer, Kappe, and Zaka, 2006)

provided a comprehensive report on some of the challenges of plagiarism detection systems. They recommended finding a means to extend the methods of authorship analysis for plagiarism detection. It was evidenced that the training of all document for a specific author based on one topic affects the performance of the classifier to detect the identity of the same author in a different topic (Koppel, Schler, and Argamon, 2009). The use of content-words alone apart from other stylistic features can help in perceiving the text content and capture the similarity with other texts to some extent. However, nowadays challenges require more than calculating the similarity between texts, the need of identifying text authorship is also essential to deal with different misconduct challenges.

3.4.2 Most Common Words (MCW)

MCW are defined as language elements without (much) inherent meaning. Their primary purpose is to clarify the relationship between words' classes in different textual parts. MCW are considered to be vital discriminating features for author's style (Stamatatos, 2009). A wide range of studies has used MCWs (function words in some contexts) which demonstrates the effectiveness of using these linguistic elements for identifying the text authorship in different settings (Chung & Pennebaker 2007; Stamatatos, 2009). The motivation behind the preference of the most common words over other word classes is the steadiness of their frequencies with different topics that were written by the same author. These characteristics of unconscious usage of MCW assist in capturing the style of a specific author regardless of the variation of subjects and minimise the risk of being deceived. In other words the frequency of MCW does not vary greatly across the topic and it is also unlikely that their frequencies can be consciously controlled (Afroz, 2013; Koppel, Schler, and Argamon, 2009).

The extending of the existing authorship analysis method based on stylometry is a promising research area that works on developing intelligent methods. These methods are relying on artificial intelligence (by involving machine learning algorithms) and stylistic analysis. Their main aim is to capture specific

stylistic patterns that can work as a unique fingerprint for each author's writing style.

3.5 **Machine Learning**

A great variety of machine learning algorithms, formulas, and techniques exist, and a detailed review of them here is beyond the scope of the present work. However, probably the most important aspect of machine-learning methods is the choice of features. A machine learning model is "only as good as the data put in", as the adage runs (Matthews & Merriam 1993). Machine learning methods often pick generously from among wide ranges of feature types, in order to generate features for their own training. In addition, they often generate features of their own such as character n-grams (described in Chapter 2), function words (of, the, to, other prepositions, etc.). In theory they dub univariate or multivariate, technique-derived features which have arisen more recently under the auspices of ML. The choice of machine learning method is also of great importance to the ultimate success of the classifier.

The nature of the learned boundaries depends on the learning method used but in any case these methods facilitate the use of classes of boundaries that extend well beyond those implicit in methods that minimise distance.

Three categories of machine learning methods are recognised; supervised, unsupervised and reinforcement learning algorithms. A supervised learning algorithm requires input from the researcher (usually a set of labels to distinguish between classes examples) for the data to be determined in order to derive the discriminant functions. During this process the machine learning algorithm reviews its response for predicting a specific class, if the response does not predict the correct label, a process to adjust the errors internally is conducted. The internal adjustments process is performed to predict the correct class labels for the future same example. The second is the unsupervised learning algorithm; where no information about classes is provided. This type of learning algorithm relies on measuring the similarities between entities and works by capturing the differences between their patterns. The third learning

algorithm is reinforcement learning which belongs to a third learning category. It is closer to supervised learning algorithms as it can receive a feedback regarding its performance from the surrounding environment. The supervised learning procedure differs from the reinforcement learning procedure because it just needs to inform if the behaviour is unfitting and how. In supervised learning precise information about the target class is given.

There are two procedures to determine the appropriate sets of features for any machine learning algorithm. The first procedure is based on selecting specific features to solve a particular problem. The second procedure includes selecting all the available features which are the so-called “brute-force” set of features. The second procedure is not useful for induction due to excess noise and therefore an efficient pre-processing procedure must be applied (Zhang et al., 2002). The literature presented the advantages and drawbacks offered by the various techniques to enhance the researchers understanding when dealing with machine learning. The use of the instances selection procedure is a good coping strategy to counter the difficulties presented when attempting to learn from very large datasets. The procedure for building an effective classifier maintains data mining quality while keeping the sample to the minimum size (Liu and Motoda, 2001). The method of a feature subset selection enables the maximum possible number of features that are either irrelevant or redundant to be determined and removed (Yu & Liu, 2004). This method also participates in reducing the features vectors dimensionality which in turn improves the machine learning algorithms.

The interdependence of many features may affect unduly the accuracy of supervised machine learning models. Construction of new features from the basic feature set is a way of addressing this problem (Markovitch & Rosenstein, 2002) and is known as feature construction/transformation. New features generated in this way can initiate more concise and accurate classifier creation, while discovery of meaningful features can make both the produced classifier and the learned concept easier to understand.

The process of deciding which machine learning algorithm to use is a critical issue. In classification procedures, the classifier works on mapping unlabelled instances to accurate classes. Evaluation of the classifier is based on three main techniques, splitting the dataset into three equal portions, one third being retained to estimate performance while the other two thirds is used for training. Cross-validation is the second technique that divided the dataset into equal size subsets with no duplication. For each subset, the amalgamation of all other subsets is used for training so that the average error rate of each subset is an estimate of the classifier's error rate. Finally, we have leave-one-out validation, which is in fact a special case of cross validation. This is more expensive in terms of computation because each test subset comprises one instance, but it does produce the most accurate classifier error rate estimate (Japkowicz & Stephen, 2002). If the error rate is insufficient the procedure is required to be reiterated. Figure 3.2 describes the process of the process of supervised machine learning algorithm and shows how the different components interact with each other.

To measure the performance of different proposed supervised learning algorithms, a specific dataset is used as experimental data and the same metrics are applied to compare between trained classifiers accuracies. Given an adequate data sample, several training sets of size N can be sampled. Then the two learning algorithms can train the training set. Calculating the differences in accuracy on a test set can be estimated for each pair of classifiers. Averaging these differences provides an estimation of the difference in generalization error that may be expected for every possible training set of size N. Variance between those estimates will indicate the classifier's variance in the total set (Nadeau and Bengio, 2003). In Figure 3.2 the supervised learning framework of is depicted and how each sub-task is performed; it describes all the sequence of supervised learning process steps. The figure starts from identifying the problem (challenge), choosing the corpus and determining the sampling procedure. Then the selection of machine learning is an important step as well, which may be chosen based on previous work or based on the problem categorisation (classification or regression). The sets of features are used as

inputs for machine learning and the training procedure is applied based on the text sampling method. Then based on the training samples that were applied to train the machine learning technique the test sample is used to measure the performance. If the prediction of the test sample is correct then the model is generated to be generalised for future external samples.

Figure 3.2. Describes the process of supervised learning, the figure was stimulated by (Zheng et al., 2006)

In the following sub-sections, two algorithms were selected as common examples of supervised machine learning techniques in authorship analysis.

3.5.1 Support Vector Machines (SVM)

Support Vector Machines are a group of correlated supervised learning algorithms used commonly for classification and regression (Boser, and Guyon , 1992; Zelenko, 2003). They are considered among the most recent, sophisticated, and high-performance algorithms in artificial intelligence. They aim to separate high-dimensional data in “hyperspace” (“space” with a dimensionality equal to the number of features derived from the training set) using a hyperplane. Although the concept of the hyperplane was invented before SVM, hyperplanes were extremely expensive to compute until the concept of the support vector was introduced. Essentially, old hyperplanes would use all of the data to derive the optimal shape for separating the data points, while support vectors only deal with the mutual k nearest neighbours or points on the border. Using only border points dramatically improves the speed and usefulness of the algorithm. The border points are further diminished, after calculation of the optimal hyperplane, to only the minimum points needed to define the plane. Also, the margin of separation is maximised. This means that the optimal hyperplane is most often the one that separates the data points between the two classes leaving the greatest possible distance between the two. In the research field of authorship analysis SVMs were broadly used and recognised for their performance on many types of problems (Moore, 2001). Abbasi and Chen (2005) stated that SVM has outperformed other algorithms such as decision tree in correctly attributing 286 cases in authorship analysis. Juola (2006) argued that the performance of the algorithm cannot be generalised to other situations where dataset and context is different.

The proposed approach is to enhance the patterns discovery using machine learning in general and SVM in particular. A significant step of preparing the training data was conducted in order to boost the classification performance.

SVM is one of machine learning techniques that was first assigned for classification tasks by Joachims (1997). Throughout the training phase, SVM builds a hyperplane that can separate cases from two different classes in the best way. The separation process is then used to decide to which side the new case will locate. The state of the art SVM has been chosen as a classifier for this research because it has successfully been used in detecting stylistic features (Diederich et al., 2003; Baroni et al., 2004). Moreover SVM is approved to be worked using different types of features; they can deal with unprocessed or pre-processed set of features.

3.5.2 Multilayer Perceptron

Neural networks were invented to imitate the human brain as they include numerous mathematical procedures to work together (Juoala, 2006). They are based on performing training by using three layers and employing a backpropagation procedure to reduce the fraction of errors between the target and the response (Rumelhart, Hinton, and Williams, 1985). Holmes, (1994) has clarified that authorship analysis based stylometric features are considered as a pattern-recognition problem. The text needs to be trained to generate models using classification procedures. The classifier is built to differentiate between negative and positive samples of a specific author's piece of text. Several researchers have used the neural networks in their studies. Merriam and Matthews (1994) have used multilayer perceptron (MLP) to discriminate between three authors writing style; dramatists, Shakespeare and Marlowe.

This section gives a brief description of MLP which can be defined as a multi-layer neural network which joins together large numbers of neurons, or units, in a pattern of connections as shown in figure 3.3. The units in a net fall into three classes: input units, receiving information for processing; output units, showing the results of processing; and hidden units which help to process the data.

In a feed-forward algorithm (Figure 3.3), signals can only travel one way, from input to output, which is first determined by training on a set of paired samples. The weights of the neurons altered to enable the algorithm classifying new

examples. For classification to happen, the signal must propagate from the input unit right through the net to determine the activation values at output units. An activation value representing a feature outside the net is attached to every input unit, this value will then send to every hidden unit connected to that input unit. Each hidden unit calculates an activation value of its own and passes the signal to output units. A simple activation function calculates each receiving unit's activation value by adding together all sending units' contributions. The contribution is defined for this purpose as the weight of the sending-receiving unit connection multiplied by the activation value of the sending unit. Further value modification usually takes place; the activation sum may be adjusted to a value between 0 and 1, or may be set to zero unless it has reached a value equal to or greater than a threshold level for the sum in question. Precise determination of the hidden layer's size can be a problem because the network can perform poorly if underestimating of the number of neurons takes place. On the other hand overfitting can be caused if an overestimation procedure for the number of nodes is assumed (Camargo & Yoneyama, 2001). Kon & Plaskota (2000) have researched the minimum number of neurons and instances that allow programming of a task into a feedforward neural network.

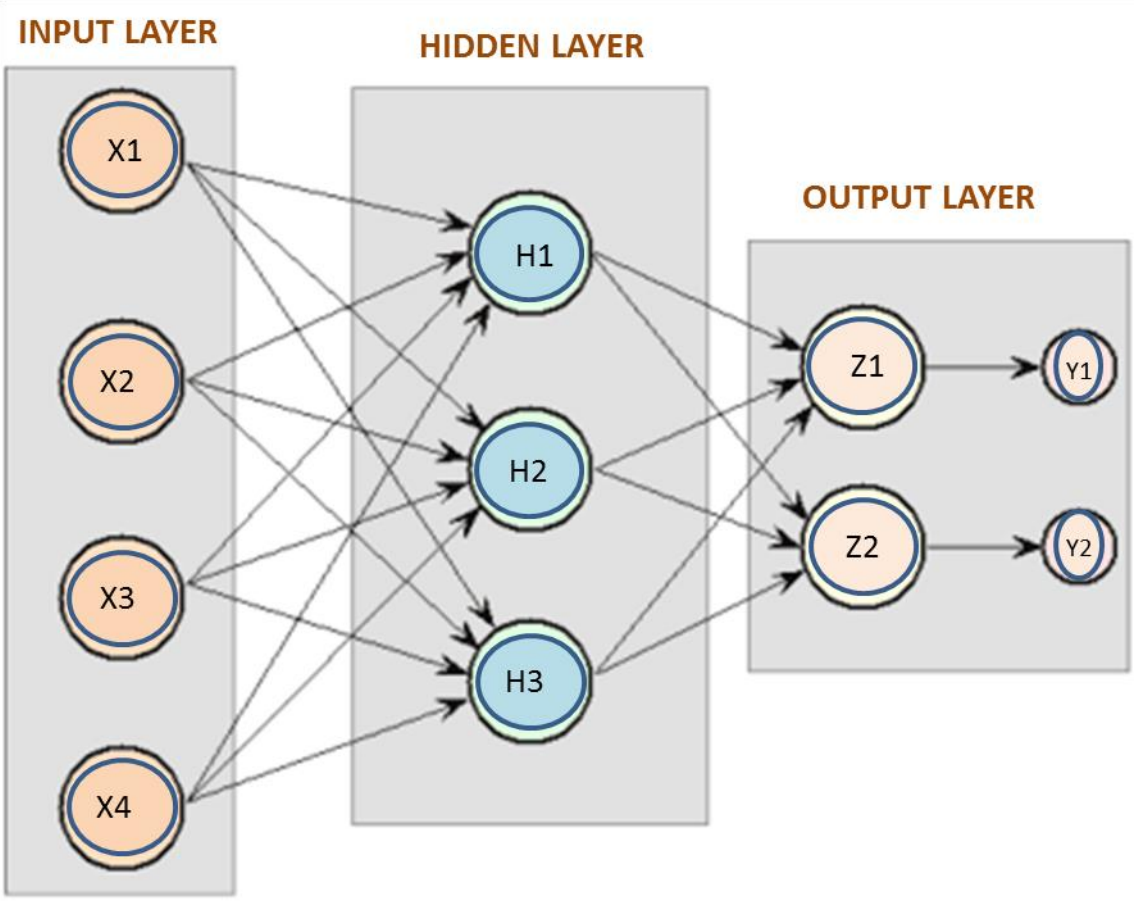

Figure 3.3: presents the three layers internal procedures for feed-forward algorithm.

There are three main aspects for any network: determining the input units, their activation functions; identifying the network structure; and then each input connection's weight. Since the first two are fixed, it is the current values of the weights that define the ANN's behaviour. At the outset, weights to be trained are given random values, following which training set instances are repetitively exposed to the net. The actual net output value is compared with the desired output value for each instance. All weights are then slightly adjusted to move net output values closer to desired output values. Neocleous and Schizas,

(2002) stated that there are many algorithms for network training existing, however the back-propagation (BP) algorithm is the most popular for estimating weight values. Kotsiantis (2007) has described the back-propagation algorithm BP using the following 6 steps:

1. A training sample is presented to the neural network.

2. The network's output is compared to the desired output for that sample and the error at each output neuron is calculated.

3. The local error is calculated for each neuron by first calculating what the output should have been and then applying a scaling factor setting the degree of upwards or downwards adjustment the output needs if it is to match the desired output.

4. Reduce the local error by adjusting each neuron's weights.

5. Strengthen the influence of neurons connected by greater weights by, in effect, assigning responsibility which known as "blame" for local error to previous level neurons.

6. Repeat each previous step on previous level neurons, this time using that neuron's "blame" as its error. The rule for updating weights is, generally speaking as explained by Kotsiantis, (2007):

$$\Delta W_{ji} = \eta\, \delta_j\, O_i$$

Where, η is defined as a learning rate and must be a positive value to clarify the size of the search step in in the gradient decent procedure. Assigning a high value to η influences the speed of BP procedure to complete the configuration of the target weight, however this also increases the risk of inability to achieve this target. O_i , is the outcome that resulted from a calculation process that was performed by the node ***i***. The value of δ_j is obtained by the following:

$$\delta_j = O_j(1 - O_j)(T_j - O_j) \qquad 3.9$$

Represents the output of the nodes and T_j represents the desired output of the node *j.*

The following formula is used to calculate the hidden nodes:

$$\delta_j = O_j(1 - O_j) \sum \delta_k W_{kj} \qquad 3.10$$

Zhang (2000) clarified that a number of weight modifications must be performed by the back propagation algorithm in order to reach a good weight configuration. Taking the number of training instances as n and the weights as W, the time taken by each repetition/epoch in the learning process is calculated as O (nW). The worst case scenario, however, gets epoch numbers as exponential to input numbers. Neural nets therefore use several stopping rules to determine the end of training. Of these, the four most common are: i) Stop when a specified number of epochs has been completed, ii) Stop when the error measure threshold is reached, iii) Stop when there has been no error rate improvement despite completion of a specified number of epochs, iv) Stop when some of the data sampled from training data shows an error measure that exceed (overfits) a pre-determined level. It is customary to train feed-forward neural networks using either the original back propagation algorithm or one of its variants. Because they are too slow for most applications, the training rate may be accelerated by estimating optimal initial weights (Yam & Chow, 2001). The Weight-elimination Algorithm is an alternative training method for multi-layered feed-forward neural networks ANNs. It derives the appropriate topology automatically, thus avoiding overfitting (Weigend et al., 1991). There has also been a use of genetic algorithms to train the weights of neural networks (Siddique and Tokhi, 2001) and to establish their architecture (Yen and Lu, 2000).

3.6 **Summary**

Plagiarism detection becomes more important recently in different life sectors. The detection process is not a straightforward procedure due to the misconduct complexities. The literature revealed that the existing techniques have many

limitations which made the detection process more difficult. It was demonstrated by the literature that bag of words (BOW), latent semantic analysis (LSA), stylometry and machine learning have a good characteristics to work on detecting plagiarism. The aforementioned methods were explored and discussed in this chapter in order to understand their performance and investigate their feasibility to be used as part of the proposed approach. BOW is considered a well-performing and starting point for generating the initial features set which produced high dimensionality in vector space. LSA is an intelligent technique that can be used for dimensionality reduction and deep text analysis to reveal the text meaning. In spite of LSA's ability to uncover the latent associations between words to capture the semantics of the text, it was not able to capture the authorial attributions. In contrast the stylometric technique performs superficial analysis based on statistical attributes for text authorship applications. LSA can reduce the dimensionality of BOW, while stylometry can enhance the performance of LSA for classification tasks. On the other hand, the most important aspect of machine-learning methods is the choice of features. A machine-learning model is "only as good as the data put in", as the adage runs. The key issue when dealing with machine learning is not to select the best algorithm, however determining the influential conditions on a given problem is more important.

The next chapter discusses the proposed extrinsic method plagiarism detection approach including its component, methods and implementation details. It also discusses how the integration between selected methods after applying some moderations can influence the overall performance.

Chapter 4: Proposed Extrinsic Method for Plagiarism Detection

4.1 Introduction

The previous chapter presents a background of the methods that were highlighted from literature and discusses their empirical foundation. It also for the purpose of this research explores the key ideas, sheds light on each method and discusses its underlying concepts.

The literature review shows that most of the automated plagiarism detection methods are based on string matching techniques (Stappenbelt, B. and Rowles, 2010; Ramnial, Panchoo and Pudaruth, 2016) and they ignore the semantic relationship between the plagiarised text and the reference document. Hence the semantics of the text is still a challenge for all supported plagiarism detection tools as discussed in chapter 2.

This chapter proposes an extrinsic method for plagiarism detection to address the gap that has been highlighted in the literature. This gap in existing extrinsic detection methods has centred on the issue of detecting text semantics and identifying its authorship. In order to measure the similarity between two texts, existing tools rely on identical text matching procedures using different algorithms. In order to address these gaps a method was proposed that includes the integration of the bag of words (BOW), Latent Semantic Analysis (LSA), Stylometry and Support Vector Machines (SVM) techniques.

The core component of this method is LSA which can be identified as an intelligent technique that analyses words co-occurrence and captures the latent associations between them in order to reveal text semantics (Deerwester et al., 1990). This method presents a new application for LSA to address the text semantics using its internal processes. It was also fine-tuned to take-in the stylometric features in order to enhance the classification procedure in

identifying the text authorship (class). Stylometry has limited function in this method as its participation has been represented by the use of the most common words. BOW was used as an initial tokenisation technique for all text. BOW tokenises text into all text words without eliminating any class of words. With this technique the books for all authors were trained according to their correct labels, the training for each book for all 292 books being performed. Then each book is tested to be classified to the correct author (class). This method was built on the assumption that each test book is compared to a collection that include books from the same author and books books from other authors. The performance of the proposed method is measured on the number of books that were classified to the correct author. SVM was employed in order to build the classifier model.

The idea of using LSA for extrinsic plagiarism detection was proposed by Ceska (2009), the study that stimulated this approach. However Ceska in his thesis was limited to applying the traditional LSA and measuring its performance for detecting text semantics. The contribution of this proposed method was to incorporate stylometry which dealt with the superficial characteristics of an author's writing style with LSA. LSA is defined as a deep text analyser that can deal with latent semantic text attribution. The integration between both of methods is to work on preparing an informative set of textual features for classification purposes.

The work proposes a novel experimental methodology for testing the performance of the proposed approach. This methodology relies upon the CEN (corpus of English novels) training datasets but divides that dataset up into training and testing datasets.

The proposed approach intends to answer the primary research question as stated in chapter 1, the research question is

How effective is the use of latent semantic analysis when combined with stylometry and machine learning techniques for the task of detecting semantic patterns and identifying which author wrote the document, when a reference collection is available for comparison?

The rest of the chapter is organised as follows: Section 4.2 describes the proposed approach, its various components and their implementation details. Section 4.3 presents a summary of the work in this chapter.

4.2 The Proposed Extrinsic method for Plagiarism Detection

This section proposes a new method to address the limitations of existing extrinsic plagiarism detection methods by developing an integrated model. The proposed method helps to extract informative features, apply deep text analyses and take advantage of stylometric analysis to enhance text identity detection.

The method is based on four renowned techniques: bag of words (BOW), Latent Semantic Analysis (LSA), Stylometry and Support Vector Machine (SVM). BOW is a robust and well-performing features generation model, it pulls all books for each author, breaking down all the works of that author into frequency counts for each word. LSA works as a feature extraction technique to shrink the high dimensional vectors space that resulted from BOW. LSA is an intelligent computational linguistic technique that offers a quantitative representation of a semantic domain, and can exploit the semantic model (Deerwester et. al. 1990). Stylometry is a statistical analysis of the text written by author to characterise the variation based on stylistic computations. SVM is a supervised learning technique used for classification task. Several components have been proposed as shown in Figure (4.1), each component has a specific computational function to accomplish the model fitting. The idea behind this method is incorporating the superficial text analysis which is performed by

stylometry and deep text analysis which is performed by LSA. Stylometry application relies on the use of most common words in order to capture the usage patterns of MCW for each author. On the other hand, LSA deals with analysing the text in order to reveal the latent associations between words. SVM works on searching boundaries between different classes. Several components were proposed, each components was assigned to perform a specific activity, however all components synchronise their activities with each other.

The proposed method is depicted in figure 4.1 consisting of 2 phases and internal 5 stages which are briefly described below.

- Phase 1 includes three main stages
    - Stage 1: In this stage the text from the corpus of English novels is organised into 25 different directories. Each directory will be named after the author, and will end up having all the books of that author. The books for each author will consist of the classes that were trained using the proposed method to recognise the semantic and stylistic features. Then an index was created of all books, by author. Table 4.2 has shown the Python script of indexing each author dataset.
    - Stage 2: Includes the application of BOW as described in section 4.2.1.1
    - Stage 3: Includes the steps of LSA fine-tuned in order to take in the most common words in its application as a boosting step for building an efficient classifier that can separate classes of authors. The fine-tuned details and application has been described in section 4.2.1.2.

- Phase 2 includes two stages;

- Stage 4: Includes the process of building the classifier by using the leave-one-out-cross-validation sampling technique. The classifier is trained on the features set that resulted from integrating BOW and LSA based stylometric features.
- Stage 5: Includes the evaluation of the proposed method performance, calculating the performance accuracy based on averaging the results of all estimators that were obtained from stage 4.

The following sections describe the main components for the proposed method together with the steps for its implementation.

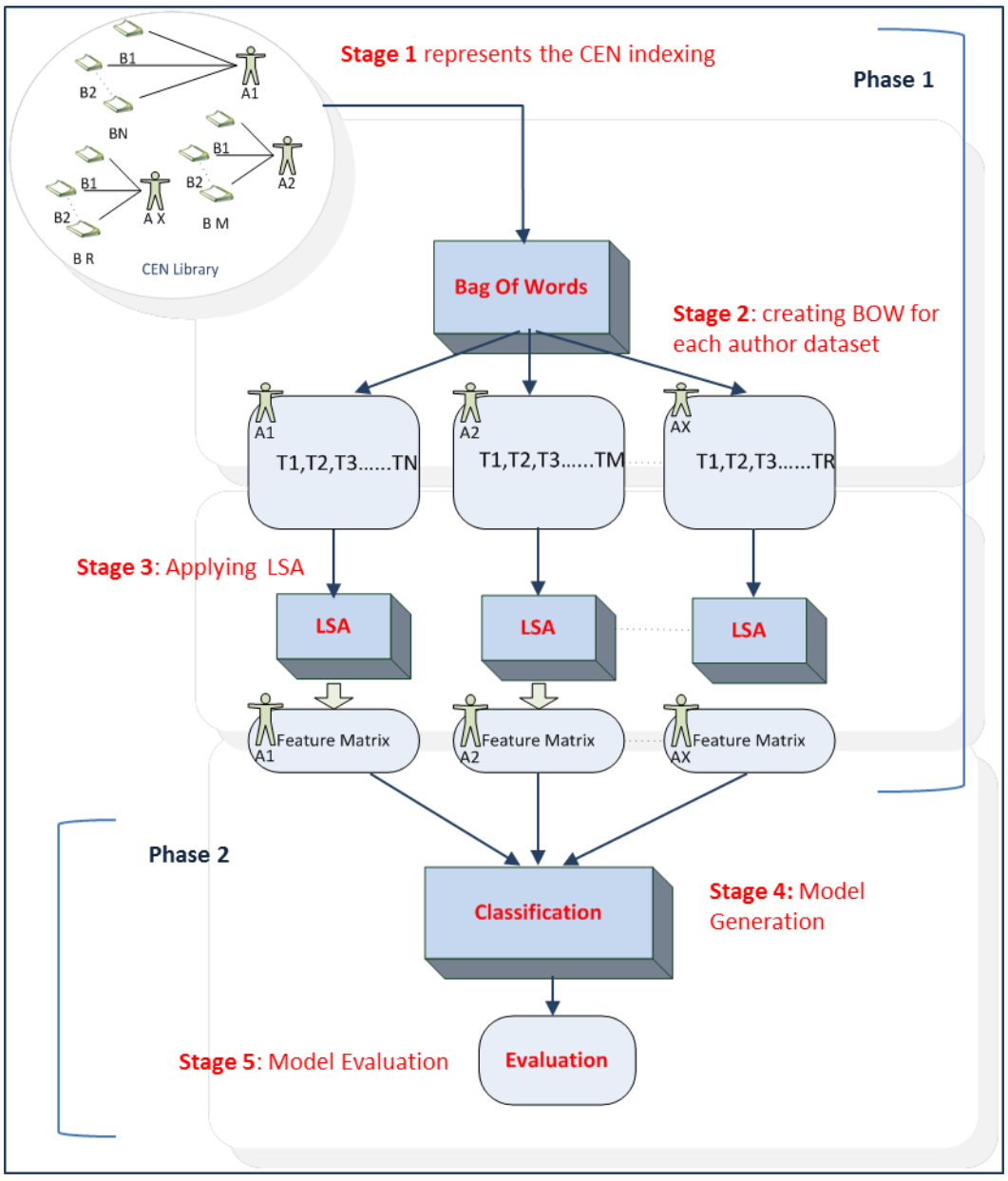

Figure 4.1. Represents the outline of the extrinsic method for plagiarism detection together with the main components; BOW, LSA and classification

4.2.1 The Components and Implementation Details of the Proposed Method

In order to perform the extrinsic method for plagiarism detection, several components were proposed to fulfil the task of capturing the text semantics and identifying the text authorship. The components and their detailed implementations are presented as follows:

4.2.1.1 *Bag of Words (BOW)*

The first component in this approach is BOW which represents the initial features generation step to represent the text. The experimental corpus is the Corpus of English Novels (CEN) which was chosen to validate the proposed approach.. CEN is composed by Hendrik De Smet, and has been used in many studies to follow the temporary changes in the text and compare its patterns usage across authors. It has been formed based on English novels, written by twenty-five novelists, British (including Irish) and North American. The novels were written in the period between 1881 and 1922; furthermore all authors were born between 1848 and 1963 and represent roughly one generation of writers. The following table 4.1 summarises the contents of the corpus.

Table 4.1. Presents a summary of the Corpus of English Novels; the authors, number of books, the publication year and the number of words

Author	NR. of Novels	Year of Publication	NR. of Words
Andy Adams (1859-1935)	5	1903-1911	450,564
Arthur Conan Doyle (1859-1930)	18	1888-1913	1,566,987
Edith Nesbit (1858-1924)	8	1899-1907	537,969
Edith Wharton (1862-1937)	11	1900-1922	872,824
Emerson Hough (1857-1923)	9	1900-1922	751,315
Frances Burnett (1849-1924)	11	1881-1922	974,948
Francis Marion Crawford (1854-1909)	13	1882-1903	1,396,223
George Augustus Moore (1852-1933)	10	1885-1901	996,682
George Gissing (1857-1903)	20	1884-1905	2,408,767
Gertrude Atherton (1857-1935)	10	1888-1922	634,864
Gilbert Parker (1862-1932)	16	1893-1921	1,398,355
Grant Allen (1848-1899)	8	1884-1899	590,205
Hall Caine (1853-1931)	4	1885-1913	665,937
Henry Rider Haggard (1856-1925)	25	1885-1910	2,556,621
Henry Seton Merriman (1862-1903)	12	1892-1913	988,647
Humphrey Ward (1851-1920)	17	1881-1916	2,252,823
Irving Bacheller (1859-1950)	8	1892-1922	511,064
Jerome Kapla Jerome (1859-1827)	10	1886-1919	706,389
Kate Douglas Wiggin (1856-1923)	14	1893-1915	677,656
Lyman Frank Baum (1856-1919)	14	1900-1916	622,700
Marie Corelli (1855-1924)	11	1886-1921	1,719,829
Ralph Connor (1860-1937)	11	1898-1921	974,840
Robert Barr (1850-1912)	10	1893-1910	731,329
Robert Louis Stevenson (1850-1894)	9	1881-1893	676,472
Stanley John Weyman (1855-1928)	6	1890-1901	563,418
Total	**292**	**1881-1922**	**26,227,428**

This dataset is represented in the initial step, using the BOW technique. BOW is a feature generation technique which splits the text into words associated with their frequencies. This method was used in a wide range of applications to represent textual documents and visualise images. BOW is used to break the documents of the corpus into unique words, counting the frequency to form the "baseline" features set.

Table 4.2: Re-organise the CEN into separated datasets

Table Script	Function

def bookdict(cen_authorship_file):
  inf = open(cen_authorship_file)
  lines = inf.read().strip().split('Corpus of English Novels\n\n')[1].strip().split('\n')
  inf.close()
  books = dict()
  for line in line publication_year = line.strip().split(',')[0].strip()
  booktitle = publication_year + ' ' +
  line.strip().split(',')[1].split('(')[0].strip()
  if "\x92" in booktitle:
  booktitle = booktitle.replace("\x92", "'")
  try:
  author_books = books[author]
  except KeyError:
  author_books = []
  author_books.append(booktitle)
  books[author] = author_books
  return books
	Re-organise all books of the corpus of English novels (CEN) by creating the initial directories.

def make_CEN_byauthor(books):
  import os
  for author in books.keys():
  # print author
  author_dir = cen_parent_dir + '\\' + author
  try:
  os.mkdir(author_dir)
  except WindowsError:
  a = 1
  author_books = books[author]
  for book in author_books:
  book = book.lower()
  if "\x92" in book:
  book = book.replace("\x92", "'")
  ## print author+', '+book
  try:
  textfile = open(cen_dir + '\\' + book + '.txt')
  text = textfile.read()
  except IOError:
  print("error")
  print(book)
  continue
  outfile = open(author_dir + '\\' + book + '.txt', 'w')
  outfile.write(text)
  outfile.close()
	CEN corpus to create one directory for each author, with all of the books for that author.

The main steps in BOW procedure are

- Create a BOW_array
- Eliminate tokens include empty string, numbers and punctuation marks.
- Tokenise the text into words splitting by spaces for each book
- Add the tokens to *BOW_array*
- Use Python_code to place the *BOW_array* in order to calculate the frequencies. All occurring words together with their frequencies are represented as shown in table 4.2 to form the row TDM matrix (term-document matrix).

Implementation of BOW

The CEN corpus was reorganised into 25 different directories according to the 25 authors using Python script. Each directory was coded by the author which will end up containing all the books for that author. The books for each author have consisted of the classes and also each file in the directory has been saved with an extension which includes the author code. In table 4.3 a script was performed in order to create the author directories and each author folder is presented.

Table 4.3 Presents the scripts of re-organising the CEN corpus and create each author's separate dataset.

Script	Function

`def bag_of_words(documents):`
`  bow = dict()`
`  for document in documents:`
`  words = document.strip().split()`
`  for word in words:`
`  try:`
`  count = bow[word]`
`  except KeyError:`
`  count = 0`
`  count += 1`
`  bow[word] = count`
`  import operator`
 	
`  order = operator.itemgetter(1)`
`  bow = [[word, bow[word]] for word in bow.keys()]`
`  bow = sorted(bow, key=order, reverse=True)`
`  return bow`
	Transform text into bag of words

`def bag_of_words_book(text_books_author):`
`  bow_author = dict()`
`  for title in text_books_author.keys():`
`  document = [text_books_author[title]]`
`  bow_book = bag_of_words(document)`
`  bow_author[title] = bow_book`
`  return bow_author`
	Creating BOW for each book

`def bag_of_words_author(text_books_author):`
`documents_author =[text_books_author[title]`
`for title in text_books_author.keys()]`
`  bow_author = bag_of_words(documents_author)`
`  return bow_author`
	Creating bag of words for each author (TDM) matrix

The BOW model was implemented using a Python script which was written to automate the extraction of each of the 25 different training sets. The pre-processing step includes retaining the words with high frequencies or MCWs, eliminating numbers, punctuation marks and blank strings. The process was performed by pulling all the books for each author, breaking them down into

frequency counts for each word. The book in each author's dataset breaks into its unique words and the frequency for each word is calculated for both MCW and content words (CW).

The application of BOW has performed the calculations of array frequencies for each book, then these books arrays are combined to establish an author's matrix. The output from BOW is the term document matrix that includes all words and their accompanying frequencies for each author (class). The following script represents the procedure for chunking the text into words; the script is recalled again in order to process each book in each author's dataset.

The fundamental data structure BOW is very high-dimensional (sparse) because documents can contain a large number of words, but share very few.

Hence LSA was proposed to manage the high dimensionality of data set resulting from bag of words

4.2.1.2 *Latent Semantic Analysis (LSA)*

An introductory step for dataset representation using BOW technique was used to tokenise the text into words accompanying with their frequencies. In order to reduce the sparseness of the term-document features that was resulted from BOW, latent semantic analysis (LSA) was applied. LSA is used as a dimensionality reduction to reduce the size of the features space from a BOW in order to make "latent" contextual clues about the meaning of words. These contextual clues are based on the co-occurrence of words in a particular context. The more compact feature space is meant to be trimmed of excess noise (random association from chance instances of co-occurrence). The goal is to achieve this compactness without losing a significant portion of information. The more manageable features matrix is also easier to run through classifiers, because it requires less computing time and can be trained efficiently. Extrinsic plagiarism detection is most similar to text-classification techniques because it compares the suspicious input document (e.g., query) to a collection of known documents. In the proposed approach, two text representatives' sets of features: most common words (MCW) and content words (CW) were proposed. MCW are defined as language elements without (much) inherent meaning. Their primary purpose is to clarify the relationship between words' classes in different textual parts and considered as discriminating features for author's style (Stamatatos, 2009). CW are artefacts of particular writing attitudes or belong to specific topics, such as nouns, verbs, adjectives, and adverbs which describe some objects, actions, or statuses. They contrast with the MCW in functions, furthermore they help to characterise text context but not author's style (Koppel, Schler and Argamon, 2009).

Itis known that MCW are associated with high frequencies, as a result, they are often eliminated during the pre-processing step in particular with LSA applications (Ceska, 2008). In the proposed approach, the main goal is to extend the authorship analysis methods which rely on MCW to characterise the authors' writing style and incorporate LSA for semantics detection. LSA was slightly modified to cope with MCW as an essential stylometric features that can help in detecting authorship and discriminate between authors (classes).

The application of LSA method includes a series of steps; pre-processing, transformation and constructing the semantic space. During the pre-processing procedure MCWs are eliminated because of their high frequency occurrences and their lack of suitability for verifying the text meaning. Even if they have been retained they will be ignored during the transformation step because they will be given trivial weight values. In order for LSA to recognise MCW as features the weighting method TF-IDF is needed to be fine-tuned. The internal procedure of implementing traditional LSA as stated above ignored the MCW. TF-IDF is the weighting method used by LSA in R. R is the computational and statistical environment used as software library to perform wide range of tasks. The weighting method TF-IDF is defined as a statistical measure used to evaluate how important a word is to a document in a collection or corpus. This importance (rank) grows correspondingly to the number of times a word appears in the document but is counterbalance by the frequency of the word in the dataset. The extrinsic plagiarism detection method is based on MCW. The method calculates the weights of MCW based on its appearance in each class rather than in a corpus for the favour of LSA application.

The TF-IDF weight value is calculated in two steps:

Term Frequency (*TF):* The first step is to calculate *TF* which measures how many times the term $\boldsymbol{T_d}$, occurs in a document. As documents are differ in their lengths, a normalisation process is performed by dividing the $\boldsymbol{f_t}$ (term frequency) by document length (the total numbers of terms in the document) $\boldsymbol{DN_t}$ as shown in equation 4.1 and 4.2

$$\text{TF}(T_d) = \frac{\text{(Number of times term t appears in a document)}}{\text{(Total number of terms in the document)}} \quad 4.1$$

$$\text{TF}(T_d) = 1 + \log(T_d) \quad 4.2$$

Inverse Document Frequency (IDF: The second step is to calculate the IDF which measures the importance of the term. The calculation of TF, considers all terms are equal in importance, but certain terms such as "is" or "of" have high frequencies without much contribution. Thus they are often either eliminated or weighed down and the rare ones scaled-up. The IDF divides the total number of documents, $\boldsymbol{TN_d}$ by The number of documents that contain $\boldsymbol{T_d}$ which is denoted by $\boldsymbol{N_{Td}}$ as shown in equation 4.3 and 4.4

$$\text{IDF}(\boldsymbol{T_d}) = \log \frac{(Total\ number\ of\ documents)}{Number\ of\ documents\ with\ term\ \boldsymbol{T_d}\ in\ it).} \quad 4.3$$

$$\text{IDF}(T_d) = \log(TN_d\ /\ N_{Td}\ ) \quad 4.4$$

The term weighting value is calculated as shown in equation 4.5

$$\text{TF-IDF} = (1 + log\ (T_d)) * log\ (TN_d\ /\ N_{Td}\ ) \quad 4.5$$

TF-IDF based on a notion that MCW that appear on one class over others can play a significant discrimination rule. MCW (terms with high frequency value) were kept and the weighting method is slightly fine-tuned to reflect statistically how important a frequent word $\boldsymbol{f_T}$ is to a document in a class $\boldsymbol{C}$ instead of corpus or dataset.

$$\text{TF-IDFc} = (1 + log\ (f_T)) * (log\ (K_c/M_t\ )) \quad 4.6$$

where, $\boldsymbol{M_t}$ represents the number of documents in class $\boldsymbol{C}$ which containing $\boldsymbol{f_T}$ and $\boldsymbol{K_c}$ represents the total number of documents in the class.

Once the weighting procedure is completed, the mathematical method; the Singular Value Decomposition (SVD) is performed to derive the model of semantic structure. The process of SVD in LSA works in the following manner.

Assuming *A* is the original matrix of document-word associations, of dimension *m* x *n*:

$$A \approx A_k = T_k S_k D_k^T \qquad 3.7$$

SVD decomposes the original matrix into three matrices in order to shrink the features space as shown in figure 4.2. The k subscript refers to the uniquely LSA contribution to the SVD operation, k is the number of singular values taken to represent the original term-document matrix, usually fairly small for efficient computation (and effective noise-reduction).

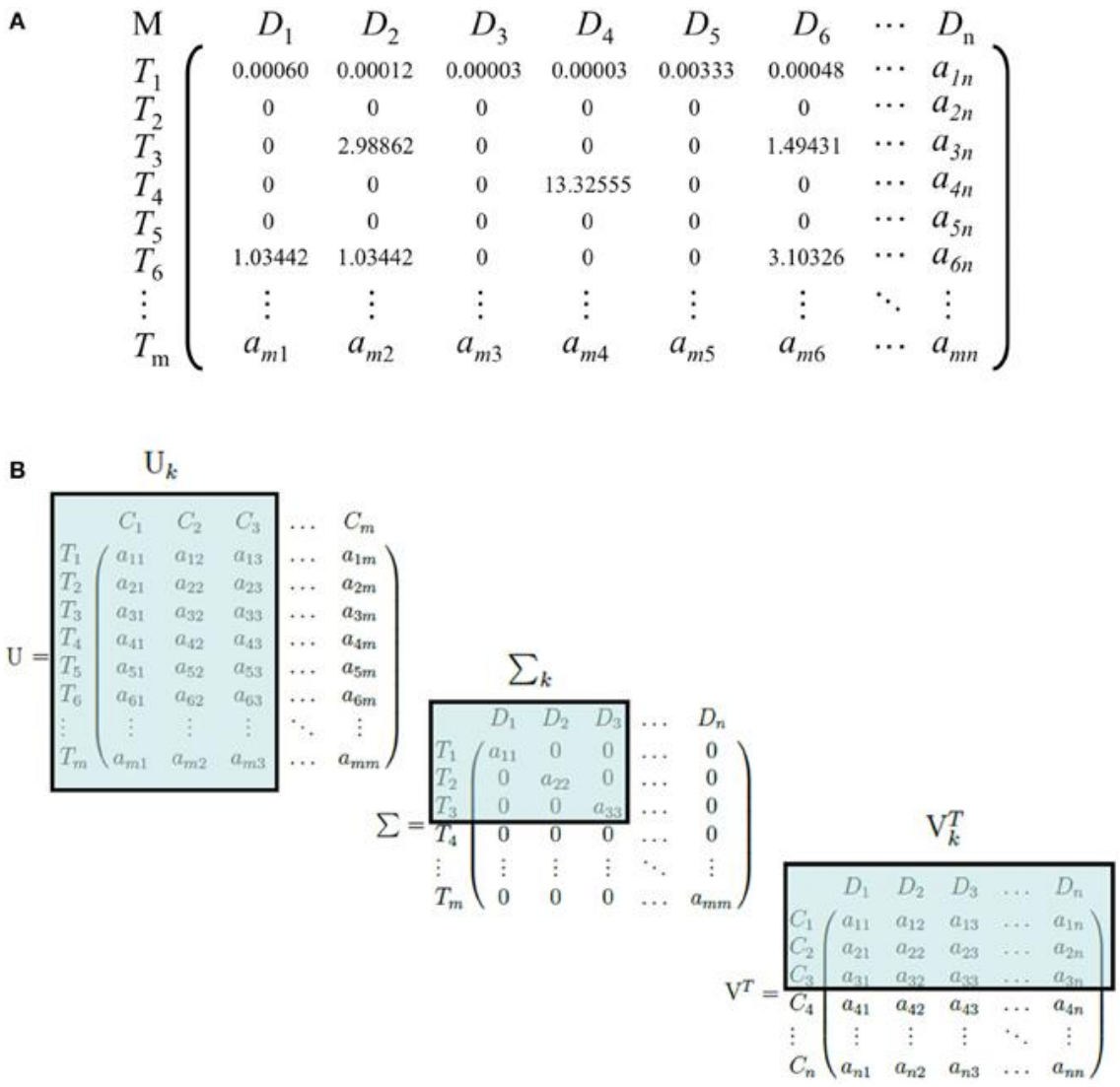

Figure 4.2: represents the process of applying SVD on the original matrix, and its effect on reducing the features space.

Implementation of LSA

Following the TDM creation by using BOG, the matrix has undergone a series of procedures based on the LSA algorithm. The raw features matrix was transformed by the normalisation technique TF which is calculated by dividing the term frequency in the document by the total number of terms in the document. The terms are re-weighted based on their appearance in classes (author's datasets). The new application of LSA in this approach is the use of most common words (words with high frequency) as an additional features set with content words CW. The use of MCW added a layer of stylometric analysis to capture the author writing style. LSA was incorporated in this approach to offer a deep linguistic analysis method that works on uncovering the latent association between terms. The Both LSA and stylometry were integrated to develop semantic models and capture the relevant patterns of MCW usage for text authorship detection.

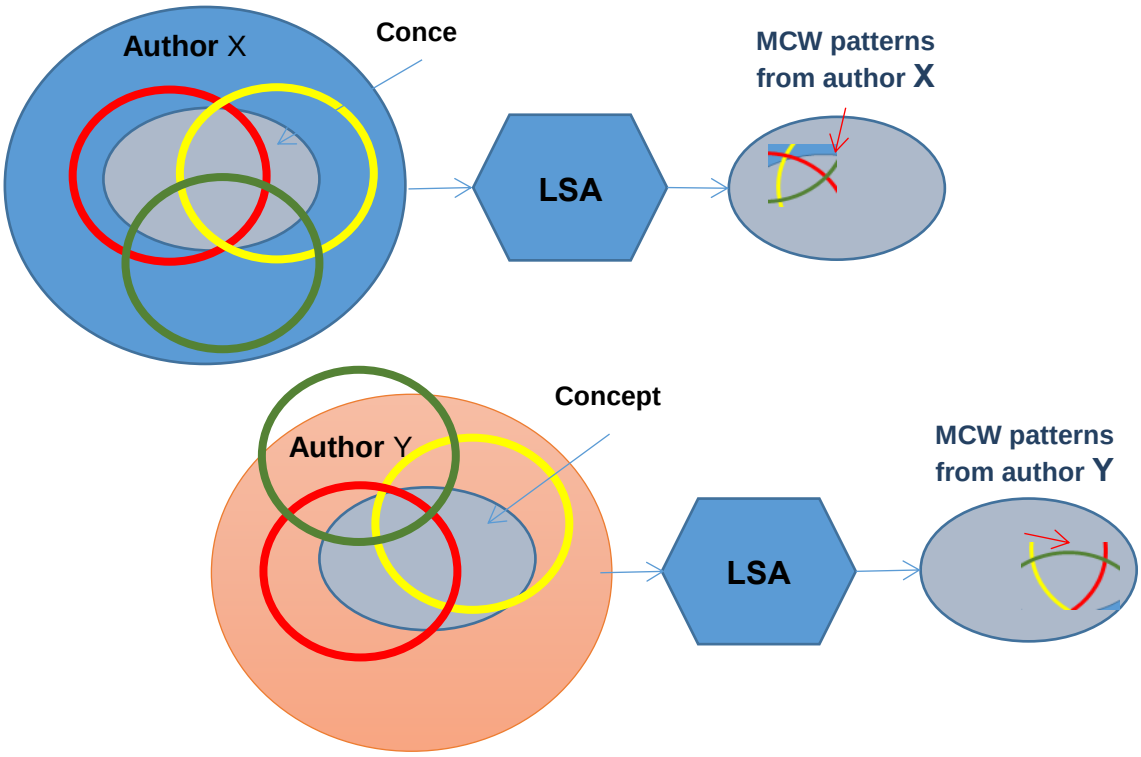

Figure 4.3: Represents the application of LSA procedure and most common words patterns capturing (source: the author)

In figure 4.3, the process of applying LSA, shrinking the space and re-weighting MCW in order to capture the stylistic is shown. The figure assumes that the coloured circles (red, yellow, green) represent different groups of MCW such as prepositions. The application of LSA leads to the shrinking of the space in order to keep the core concept of the text and the same time capture the MCW usage patterns.

In the process of constructing the semantic space, the features were re-weighted and a layer of stylometry has been constructed by adding MCW as stylistic features. The core algorithm of LSA, SVD is applied to create the semantic space.. The matrix which was re-weighted by enhancing the MCWs weights (as clarifies above) was assumed to be as a proactive discriminative step. It aims to discriminate between the MCWs usage patterns of the author in order to improve classifier performance. The SVD decomposed the re-weighted matrix into three special form matrices as shown in figure 4.2. The new matrices break down the original word connections and re-form an approximation relationship with closer concepts under fewer dimensions. By applying the dimensional reduction technique SVD (to experimental k), so the meaning or semantic correlations can be expressed within the new shrink space by k factors. The k value was set to 5 in all applications. The k subscript refers to the LSA contribution to the SVD operation that differs from other dimensionality reduction techniques. The value of k is the number of singular values taken to represent the original term-document matrix, usually fairly small for efficient computation. The pre-processing procedure has relied on removing punctuation marks and special strings.

The module LSA ("Latent Semantic Analysis") by Dr. Fridolin Wild written in **R** was used (Wild, and Stahl, 2007). The package is open source and can be found in the Comprehensive R Archive Network (CRAN), this package offers a high level of abstraction to facilitate the using of Latent semantic analysis. Wild offers the function "***textmatrix ( )***" to call documents from their directories, then the documents are converted into TDM matrices. With the ***lsa ( )*** function the Singular Value Decomposition SVD is applied to construct the semantic space

as shown in figure 3.5. Furthermore the package also offers several tuning procedures and allows a wide range of options to modify the core routines and various support connections that can assist in setting the parameters. The following algorithm describes the main steps in the application of LSA using **R**.

Algorithm 4.1 LSA on R

```
1 library (“lsa”) #load package
2 #load training texts
3 trm = textmatrix (“trainingtexts /”)
4 trm = lw _bintf(trm) * gw _idf(trm) #weightning
5 space = lsa(trm) #create LSA space
6 #fold-in test and training-books
7 tem =text matrix(“books/”, author =rownames(trm))
8 tem =lw _bintf(tem) * gw _idf(tem) #weightning
9 tem_red =fold_in(tem, space)
```

4.2.1.3 *Classification*

As described above that LSA was used as a feature extraction method in order to enhance the classification process. SVM is a featured classification technique in machine learning developed for the binary classification task (Cortes & Vapnik, 1995). The training procedure is performed by training a set of examples, each of them was then labelled to a particular class (author). Then SVM learning algorithm generates a model that assigns new examples to one of the labelled classes. An SVM model represents the training examples as space’s points which are then mapped into two separated classes. The margin between two classes (as shown in figure 4.4) needs to be clear and as wide as possible. The test example is then mapped into the SVM space and need to be classified to one of two classes on both sides of the hyperplane. The SVM only deal with the mutual k nearest neighbours or points on the border. The border points are further diminished, after calculation of the optimal hyperplane, to include only the minimum points needed to define the plane.

Figure 4.4. Selection of the optimal hyperplane depends on a tradeoff between maximizing the margin of separation and minimizing the number of points, although the former is typically prioritised (Vapnik, 1993).

The SVM can use several kernels, such as the radial basis function, linear, or polynomial. The kernel function aims to divert the objects into a higher dimensional space as a proactive step in order to apply the classes' separation in non-linear space. The SVM attempts were applied using three methods. the first is the linear SVM application. The second is SVM with radial basis function (RBF) kernel, which is used to make a nonlinear feature map (implicit). RBF works on adding a"bump" around each data point. The third is the implementation of the sequential minimal optimization algorithm (SMO) which was used to perform the training process for support vector classifier SVM (Platt, 1998). This algorithm can replace all missing attributes and efficiently transform the nominal ones. It also works on normalising all attributes and

provides the output based on normalised attributes. However, this algorithm can take a longer time to perform training for some attributes.

Implementation of SVM for classification

This component is responsible for building the model by applying SVM for classification purposes. The extracted feature values that resulted from the LSA application as described above are used for classification purposes. Briefly the procedure starts by training the SVM algorithm in order to construct the extrinsic plagiarism detection model. Then the model is employed to predict the accurate class from the test examples. Spontaneously, the training samples are represented as points in a particular space by the SVM model to search for boundaries between two classes. The points which represented the classes are separated and the gap between the two separated areas needs to be kept as wide as possible. The test sample is then projected into the same space to be predicted by the model in order to decide to which class the new sample belongs.

The ***LIBSVM*** **has been used from** ***e1071*** package with **R** interface; the independent variable (y) was assigned to switch to classification mode in the R engine (Chang, and Lin, 2011). LIBSVM applies one *vs* classification method to obtain the cross validation accuracy so that the highest accuracy is then returned. The SVM with RBF kernel depends on two parameters **(**C **and** γ) with the tune.svm () function. For this method the recommendations by Chen et al. (2005) have been followed, they set the same value for both parameter selections (C, γ) for all classifiers.

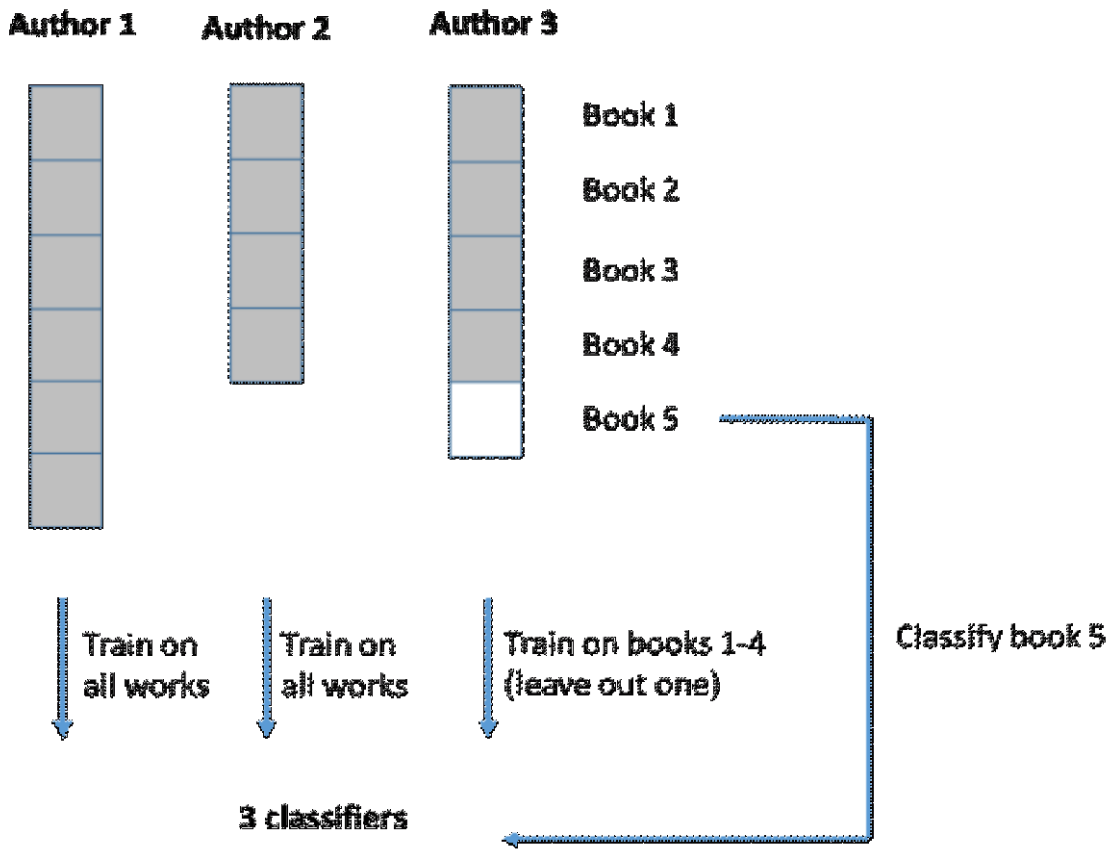

Figure 4.5: Leave-one-work-out extrinsic validation of LSA based stylometry plagiarism-detection methodology. The training performed using SVM, the numbers of authors and books are used as examples

The Sequential Minimal Optimization (SMO) learning algorithm was proposed as a learning algorithm. The core idea of SMO, during the learning process when it adjusts the weight of one example, it also adjust the weight of another. In addition SMO-SVM learning method trends to learn frequent, infrequent and unusual patterns during the application of the learning algorithm in an eager way. To apply SMO a few steps need to be considered, after selecting SVM, some parameters needs to be adjusted according to the task, context and data set. In addition some parameters can positively influence the performance of the proposed approach.

4.2.2 The Model Evaluation

Cross-Validation (CV) was used to evaluate each estimator's performance on the full set of data, demonstrating how well the algorithm performed on test data. Leave-one-out-cross-validation (LOOVC) evaluation method was applied in order to generate unbiased results. Cross-Validation (CV) was used to evaluate each estimator's performance on the full set of investigational data, demonstrating how well the algorithm performed on test data. Leave-one-out-cross-validation (LOOVC) evaluation method was applied in order to generate unbiased results as shown in figure 4.5. This method partitions the data into 292 books as a training sample. Then out of 292 books, one book is left for validating the classifier. The remaining 291 books are employed for training the classifier. The leave-one-out-cross-validation procedure is then repeated 292 times and 292 results from the folds are calculated to be averaged in order to produce a single estimation. This sampling method was chosen to better cope with detailed learning for each class characteristic.

The process of LOOCV has iterated 292 times and the results from 292 tested samples are calculated, then the results of all 292 books are averaged to produce a single estimation. The detailed results are presented in chapter 6 as the performance of this approach was measured using several metrics. The algorithm 4.2 has described the sampling method procedure that was used to measure the performance of this approach.

Algorithm 4.2 Leave-One-Out-Cross-Validation procedure.

1. The dataset is divided into k examples.
2. One book out of the k is left for validation purposes
3. The residual k-1 books are used to train the model.
4. The process is repeated n times, with each trial set used once as validation data.
5. The n results are averaged to provide a mean error over the individual cross-validation processes.

4.3 **Summary**

This chapter has presented a novel extrinsic method for plagiarism detection. The approach is based on the integration of four well-known algorithms namely bag of words (BOW), latent semantic analysis (LSA), stylometry and support vector machine (SVM). BOW breaks a document into all of its unique words; and into two types of elements content words (CW) and content-free, most common words (MCW) then calculates the frequency of each word. LSA is the core component in this approach which offers a new application when incorporated with stylometry to offer a superficial analysis for the text. It works oncapturing the usage patterns of the stylistic features; most common words (MCW).. and also uncovers the latent associations between words to captures semantics. However it is known that LSA neglects the MCW in many applications due to their high frequencies in the text. They also have no contribution to the process of revealing text meaning as they are considered as content free linguistic elements. To overcome this limitation, LSA was fine-tuned to take-in the MCW by boosting their weighting values. The MCW for each author were harnessed in the method to be captured as stylistic attributes.

SVM was combined with BOW, LSA based stylometric features to build a detection classifier. The corpus of English Novels CEN is divided into 25 datasets. Each author represents a separate class which includes all the author's books. The experimental methodology allows for a level of cross validation, the leave-one-out-cross-validation (LOOCV) method was performed

at the level of individual books, the level provides a bar or metric of extrinsic performance, as each book classified already has some other books by the same author as examples.

The next chapter discusses the proposed intrinsic method for plagiarism detection together with its components and implementation details. The intrinsic method for plagiarism detection differs from the extrinsic method which was presented earlier in this chapter. The extrinsic method for plagiarism detection uses no direct comparison to a reference collection. The key purpose this method is its use to quantify changes in writing style in order to model the ability of humans to detect variations in writing style, by training one author documents.

Chapter 5: Intrinsic for Plagiarism Detection

5.1 Introduction

The previous chapter has presented an integrated extrinsic method for plagiarism detection which deals with plagiarism semantically when the reference collection is available for comparison. This chapter presents a new intrinsic method for plagiarism detection in order to identify the text authorship. Stein and Eissen (2007) have discussed the case of how to detect plagiarism if the references are not digitised. Intrinsic methods for plagiarism detection use no direct comparison to an external reference collection. Instead, they use stylometric analysis to generate the characteristic features of an author's (i.e. a suspected plagiarizer's) body of work. The new writing output from an author is compared to other examples of the author's work in order to capture variation in the writing style (zu Eissen, and Stein, 2007). This approach is entirely different from the previous one as it is based on a set of content-free features which are the frequencies of MCW. The Intrinsic method for plagiarism detection is known as the authorship verification method which is related to forensic analysis. These methods rely on analysis of the writing style in order to quantify some stylistic features. These calculations are then employed to capture the variation between the suspicious text and others from the same author. This type of method is targeting the authorial attributes and ignoring any content or related topic. The core component of this method is stylometry, which relies on deriving sets of features based on MCW frequencies. The performance of this method will be measured based on how the derived sets of features perform with different machine learning algorithms. The use of other components is limited to the pre-processing stage. BOW has used as a first step features set generation that includes just MCW, LSA was used as the means of shrinking the vector's dimensions.

Most linguistics experts have argued that each author has a specific group of MCWs that feature their writing style and are assumed to be unchangeable (Smith and Witten, 1993). Two main factors support the importance of MCWs; their unconscious usage and corpus independence. It is not possible to find an author who writes different documents using the MCWs in different patterns such as the frequency of words (Boukhaled and Ganascia, 2015). Intrinsic methods for plagiarism detection (or authorship verification) have their roots in a linguistic research field called stylometry (Zheng et al., 2006). These types of methods are trained to infer stylistic variation in the written text of the author (Stein, Lipka, and Prettenhofer, 2011).

This chapter presents the proposed intrinsic method for plagiarism detection which like the extrinsic method also relies on the integration of BOW, LSA, Stylometry and machine learning techniques except that they are applied differently. The method was tested using the CEN (corpus of English novels) training datasets. The data set was divided into 25 datasets (authors books). Each book was chunked into the most common words (MCW) that appear in each book in the dataset.

The proposed method is intendingto answer the primary research question as stated in chapter 1:

How effective is the use of machine learning approaches based on most common words frequencies and their derivatives for the task of detecting stylistic patterns in the text in order to verify the target author, when a reference collection is not available for comparison?

The intrinsic method for plagiarism detection method is developed and evaluated using the corpus of English novels (CEN) as described in Section 4.2.1.1. This chapter is structured as follows: Section 5.2 describes briefly the proposed approach, its components and implementation. Section 5.3 presents a summary of the work in this chapter.

5.2 The Proposed Approach

The proposed intrinsic method for plagiarism detection is based on four well-known models namely, Bag of Words (BOW), Latent Semantic Analysis (LSA), Stylometry and multilayer perceptron.

This approach has two main phases and each phase has several steps:

Phase 1: This phase deals with text representation and features preparation for the second phase, this phase has three main steps:

a. Creating bag of words by using just the most common words as content-free features.

b. Applying LSA to shrink the high dimensional vectors space that resulted from BOW. LSA application is limited to work as a dimensionality reduction mean.

c. Deriving the proposed features set from the MCW frequencies.

Step (b) and (c) forms the features engineering (FE) component as will be described in details in the following sections.

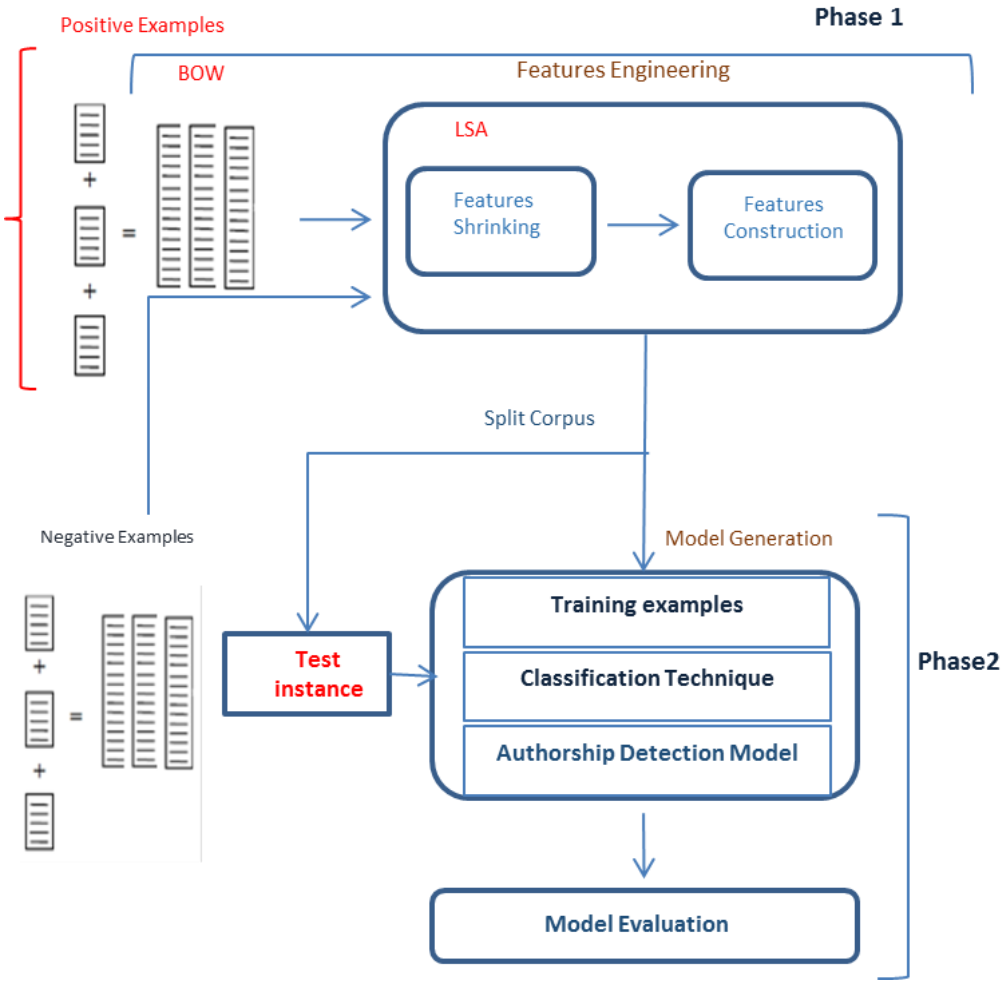

Figure 5.1. Machine Learning Model for Intrinsic Validation using Stylometry and LSA

Phase2: This phase includes also different steps:

a. Applying LOOCV method as a data sampling method for training just the target author books as positive examples in one-class technique.
b. Applying MLP as a classification algorithm to train the positive examples of the target label books based on derived features sets.
c. Then a text example is used to generate the prediction model.

The two phases of this approach are performed using several components, each component works on roles that are differ from the previous chapter. This approach has relied on the derivatives of MCW frequencies to build the prediction of authorship model. The four proposed sets of statistical features are derived in order to capture each author's writing styles patterns:

- The frequencies of the MCWs;
- The relative frequencies of each MCWs;
- The in-series proportional frequencies (e.g. 2nd / 1st, etc.); and
- The z-scores; a statistical measurement that count the number of standard deviations above and below the mean.

The third and fourth sets of features mainly rely on estimating the probability from adjacent words. This kind of estimation has played a strong role in disclosing important connections between MCWs and exposing their usage patterns. A number of different machine learning models were generated, and trained using the features described above. It was desired that a representative of each major class of machine learning methods be used. Therefore, a neural network (multilayer perceptron, or MLP), a Bayesian network (BN), a support vector machine (SVM), and a random forest (RF) were all generated, one for each of the 4 word frequency schemes.

With regard to the application of the proposed techniques one-class classification was employed to conclude if the test document was written by the target author (the author that was trained). The specific author's documents are the positive training, while the negative are anonymous without specific labels. Leave one-out-cross-validation (LOOCV) is used in order to switch the roles between texting and training data for full insights prediction performance evaluation. The algorithm of LOOCV was described in chapter 5.

5.2.1 The Components and Implementation Details of the Proposed Approach

The following subsections discuss the main components and their implementations in the intrinsic method of plagiarism detection. The reader may find the components are similar to the previously explained extrinsic method for plagiarism detection; however their procedures and implementations are different. Bag of Words

5.2.1.1 *Bag of Words (BOW)*

As we stated in the previous chapter the BOW method breaks a document into all of its unique words and counting the frequency of each word. In the field of stylometry the most common words are salient elements. This approach is entirely reliant on the most common words frequencies. No content words or any other linguistic elements were used, except a set of MCW frequencies which were used to represent all the books. The bag of words which created for this approach includes just MCW and their frequencies. A detailed description about the bag of words BOW paradigm was in the previous chapter.

Implementation of BOW

The initial text representation step using the Bag of Words (BOW) model, breaks documents into MCWS words, counting the frequency of each word, forming the baseline for each author's documents. Each document in the CEN dataset which contained sets of books for 25 authors was represented by BOW, so each author had several BOWs based on the number of documents in the author's dataset. The following points clarify the first phase procedure

1. The MCW frequencies are calculated, and then each book has its own list.

2. After the frequencies calculation for all books, the (N x M) matrix where N is the number of books and M is the number of MCWs, is constructed. The frequency matrix can be expressed as follows.

$$Fr = \begin{pmatrix} fr_1 & \cdots & fr_{1N} \\ \vdots & \ddots & \vdots \\ fr_{M1} & \cdots & fr_{MN} \end{pmatrix} \quad 5.1$$

Each column represents the most common words in each book, and each row represents the distribution of a specific word in all books which can be denoted by $\boldsymbol{f_{w/b}}$.

The BOW method has been used to generate different initial features set using just content-free words. In the previous chapter both MCW and CW were generated in order to capture semantic and authorial attributes. This method works on capturing the usage of writing stylistic features to identify the text authorship without the use of content features.

The Python script creates a specific BOW including only the MCW for each author. This represents phase (1) as pointed out in figure 5.1.

Table 5.1 presents the script of creating BOW using just common words

The script of creating the bag of words based on most common words

```
def calculate_most_common_word(bow):
    freq_dict = dict()
    for author in bow.keys():
        author_bow = bow[author]
        for title in author_bow.keys():
            book_bow = author_bow[title]
            for word in book_bow:
                word, f = word
                try:
                    existing = freq_dict[word]
                except KeyError:
                    existing = 0
                existing += f
                freq_dict[word] = existing
    import operator

    order = operator.itemgetter(1)
    freqs = [[word, freq_dict[word]] for word in freq_dict.keys()]
    freqs = sorted(freqs, key=order, reverse=True)
    return freqs
```

5.2.1.2 *Features Engineering (FE)*

This component includes two sub-components; features shrinking and features constructing. Itss known that most common words have high frequencies in the text, however authors have their special set of MCW. To reduce the vectors space of MCW that resulted from BOW, LSA has been used as features space shrinking. The goal of using LSA to capture the usage patterns of MCW for each author. Another sub-component is features construction. Researchers have demonstrated the ability of LSA in capturing the transitivity relationship between features' words. An example of transitivity (order co-occurrence) is shown in Figure 5.2 which simplifies the transitivity correlations using Doc2,

Doc2 and Doc3 as an example. Doc1 contains word A and word B and connects with Doc2 by word B that occurs in both documents. Doc2 in turn connects with Doc3 by word C, as a result word A connects to word C by word B, as a sequence word A connects to word E as a 3rd co-occurrence level. After text transformation using BOW and LSA, another pre classification step was applied and feature sets based on MCW were defined.

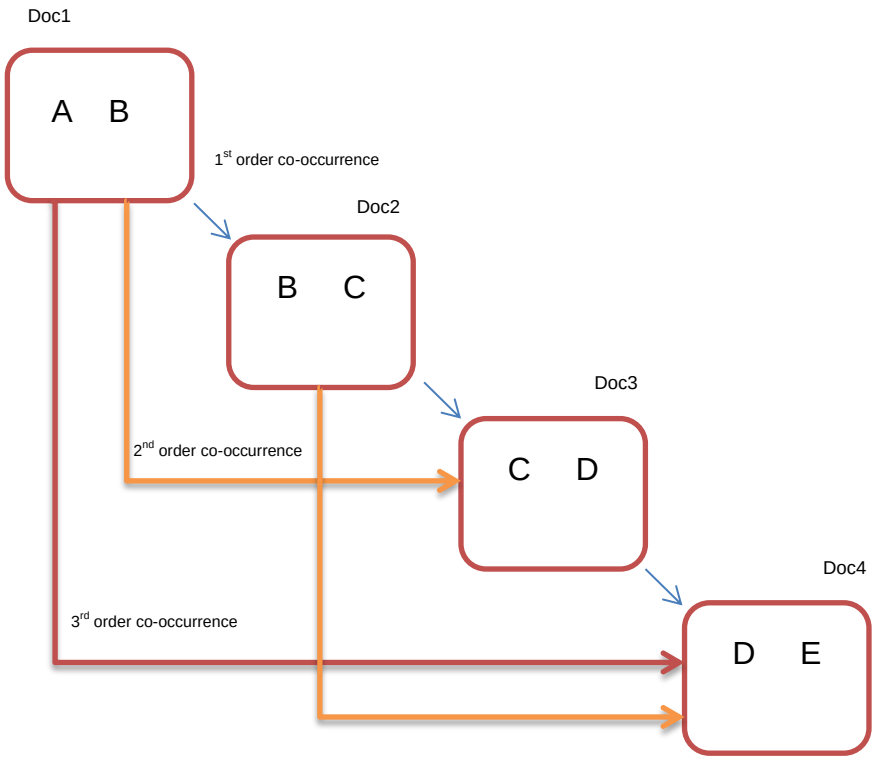

Figure 5.2. An example of the order co-occurrence tracing (source: author)

The process of feature selection is assumed to be a key issue in most applications including authorship analysis. It was evidenced that relying on the content words to detect the similarity was easier than detecting the author identity (Zhao and Zobel, 2005).. In order to explore the landscape of all authors the top ten MCWs for all authors were calculated by pooling the novels for all the authors together. Table 5.2, represents a sample of these calculations

depicting the variation of MCWs usage between different authors which determine how the quantities of MCWs vary from author to author.

The calculation results for the most frequent word varied by author, although every author had “the” as the MCW (see Table 5.2, second column). This unanimity in MCW appears in the commonness of the word “the”, as it was approximately twice as common as the second-most common word, for all authors (Table 5.1, second and third columns). It is clear that the MCWs usage is varied from author to author as for example George Gissing (the, 83943) while Kate Douglas (the, 19490) as shown in the sample table 5.2.

Another interesting note, the counts for succeeding words appeared to be closer and closer. For example, for Francis Marion (top row, below header), the counts for the 9th and 10th MCWs only differed by 47, whereas for the same author, the counts for the 1st and 2nd MCWs differed by nearly 27,000. This trend is uniform across all authors, and explains why the ranking of words varies so much across authors. The ranking of words becomes more variable, as the rank descends. For the 7th most common word, for example, the words “he”, “was”, “her”, “I”, “in” are visible across authors, whereas for the 3rd MCWs only “of”, “to”, and “and” are visible.

Table 5.2. The word, followed by the number of instances in which it occurs without punctuation, is shown. The top ten most common words for all authors were calculated by pooling all novels by all authors together

Author	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Francis Marion	the, 67793	and, 40902	of, 34522	to, 34375	a, 25921	in, 21589	he, 20864	was, 19037	that, 17015	his, 16668
Robert Barr	the, 41535	to, 19547	of, 18212	and, 15785	a, 14225	I, 11418	in, 10087	you, 9228	that, 8763	he, 7907
Henry Seton	the, 44863	of, 22590	to, 20746	a, 19868	and, 19121	in, 13512	was, 12652	he, 10598	his, 9840	that, 9496
George Augustus	the, 42723	and, 25643	to, 24636	of, 22393	a, 16755	I, 12731	in, 12697	was, 11448	that, 10587	she, 10110
Ralph Connor	the, 50022	and, 27446	to, 23024	of, 23017	a, 19686	in, 14159	his, 13723	he, 11645	I, 10881	was, 10814
Emerson Hough	the, 34685	of, 20462	and, 15590	to, 14540	a, 11723	I, 10905	in, 9335	was, 7025	that, 6340	as, 5529
Jerome Kapla	the, 32615	to, 20022	and, 18828	of, 17084	a, 16678	I, 12570	in, 8843	that, 7861	was, 7596	it, 6892
Frances Burnett	the, 35172	and, 30466	to, 25954	of, 21834	a, 21255	was, 16295	he, 14304	in, 13717	had, 12971	she, 12360
George Gissing	the, 83943	to, 63833	of, 62002	a, 50655	and, 45151	I, 32429	in, 32101	was, 29095	he, 25247	her, 25113
Kate Douglas	the, 19490	and, 12875	of, 9265	to, 8895	a, 8577	in, 6022	was, 4579	her, 4083	I, 3896	that, 3802

Based on the above investigation, the features for this proposed approach have been selected based on the baseline frequencies. Once these baseline frequencies were calculated, their frequencies were expressed as a proportion of the total word count in the book. Then, the words were re-expressed as proportions in-series, so that an in-series proportional frequency was represented by (2nd / 1st, etc.). Finally, the z-value of the variance respective to the overall set of documents for the author, are all calculated, for each book.

The proposed approach has used z-scores as a mean for predicting significant changes in MCWs usage. The equation 5.4 shows the calculation procedure of z-score by using the mean and standard deviation formulas. The importance of z-score is to show the abnormality behaviour of most common word between the test book and other books of the target author. To clarify, if zw/b < -1 the MCW appear less frequently in the tested book than its usually distribution in other books in this class than other classes.. Furthermore if zw/b > 1 means the MCW appears more frequently in this class over other classes. The idea is to measure the distribution of each class of books. The significance of the above metric is supported by the notion that analysing internal text structure can enhance the process of capturing variance patterns. Due to the specificity of the intrinsic method for detecting plagiarism sets of features that rely on the raw frequencies of most common words were proposed.

Implementation of FE

As we clarified above an innovative feature engineering method was developed and applied in two steps.

Step 1: Co-occurrence feature extraction using LSA

Latent Semantic Analysis (LSA) was employed to extract the co-occurrence feature matrix from each author dataset (25 datasets) and reveal the author's specific patterns based on MCWs. It also directly models the relationship between MCW on the basis of the usage they share. The application of LSA results in the construction of the statistics features matrix. To perform this procedure the module "LatentSemanticAnalysis" based on Weka was proposed

(Hull, 2009). The software libraries offers different types of LSA application such as the statistical method, concrete method and instance method each module having different functionality. As the goal is to explore the landscape of MCWs across different authors, so the straightforward static method which deals with space shrinks and returns the attribute transformer with a rank approximation value = 0.75.

Step 2: Feature construction

The second step is features construction, where statistical groups of features have been devised including: frequencies (as a proportion of total words), in-series proportional frequencies (2nd / 1st, etc.), and the z-value of the variance respective to the overall set of documents for each author. The z-score ($\boldsymbol{z_{w/b}}$ is calculated using the mean $\boldsymbol{\mu_w}$ and standard deviation $\boldsymbol{\sigma_w}$**).**

$$\mu_w = \frac{1}{N}\sum_{b=1}^{N} fr_{w/b} \quad 5.2$$

$$\sigma_w = \sqrt{\frac{1}{N-1}}\sum_{b=1}^{N}(f\, r_{w/b}\text{-}\mu_w)^2 \quad 5.3$$

$$z_{w/b}=\frac{fr_{w/b}-\mu_w}{\sigma_w} \quad 5.4$$

Such calculations express the deviation of the MCW frequency in each book when compared to the corpus average.

Table 5.3 Presents the script codes of each function in features construction step

Function	Script

The calculation of the words are re-expressed as proportions in-series for each **book**	`def calculate_series_freq_prop_book(bowprops):`
	`  sprops = [bowprops[i][1] / bowprops[i - 1][1] for i in range(1, 11)]`
	`  return sprops`

The calculation of the words are re-expressed as proportions in-series for each **author**	`def calculate_series_freq_prop_author(bow):`
	`  sprops_author = dict()`
	`  titles = bow.keys()`
	`  for title in titles:`
	`  bookbow = bow[title]`
	`  bookbowprop = bag_of_words_byprop(bookbow)`
	`  sprops_book = calculate_series_freq_prop_book(bookbowprop)`
	`  sprops_author[title] = sprops_book`
	`  return sprops_author`

The calculation of the average for z-score-calculation	`def calc_avg(numlist):`
	`  avg = sum(numlist) / len(numlist)`
	`  return avg`

The calculation of standard deviation for z-score calculation	`def stdev(numlist):`
	`  avg = calc_avg(numlist)`
	`  var = sum([(num - avg) ** 2 for num in numlist])`
	`  std = (var / (len(numlist) - 1)) ** (float(1) / 2)`
	`  return std`

The calculation of z-score	`def calculate_variance_props_externalauthor(author_stats, external_props):`
	`  authorprops_avg = author_spi_stats[0]`
	`  authorprops_std = author_spi_stats[1]`
	`  zvals = []`
	`  for i, ext_prop in enumerate(external_props):`
	`  var = ext_prop - authorprops_avg[i]`
	`  z = var / authorprops_std[i]`
	`  zvals.append(z)`
	`  return zvals`

In other words, the z-score means the process of measuring the abnormality behaviour of MCW frequency with regard to the corpus statistics. In order for these features to be computed, the python script as shown below in table 5.3 was written.

5.2.1.3 *Authorship Generation model (AGM)*

The component includes three procedures; training the examples, applying the classification (multilayer perceptron) then generating the model. The classification procedure in the case of intrinsic detection was applied differently as no reference documents are available in this case.

The traditional classification process includes documents with labels, each label belonging to a specific class. For this research a method was proposed to facilitate the classification process for an individual class of documents (Roberts et al., 1994; Koch et al., 1995). A set of documents was labelled to a specific classwhich was named positive examples. Another action also needs to be done to enhance the classification process in order to generate outlying samples which are named as negative examples. The positive samples represent the class of the target author and all the samples from this class are trained. For the negative examples, training is performed for all other author samples without identifying the labels of the classes, so all other author's books are labelled to be negative. Figure 5.3 describes the mechanism for the proposed classification procedure that was used in intrinsic method for plagiarism detection. The training method of this method differs from the extrinsic method as just one-class examples need to be trained.

Figure 5.3. Describes the classification method that adopted in the proposed intrinsic method, as shown in the figure (Tax, 2001).

Implementation

The one-class classification procedure is implemented in order to train separately the data description for each class which is called the target class. The features sets of the target class are considered as positive examples and all the data for the other classes are considered negatives (outlier data). The training procedure was performed on the target author documents that were already labelled by their names, so there is just one labelled documents to be trained. In order to apply an efficient classification process there is a need for another type of information just to balance the classification. This information is called negative examples and they do not need to have a specific label. The classifier needs to decide if this text belongs to the target author (tampered-free) or not (tampered).

It is worthy of note that T represents the unique source of quantifying the author's stylistic features which will be used to build a model that can track variation in the writing style. The other part which is named the negative examples are used to generate the abnormality against the target class as depicted in figure 5.1.

A multilayer perceptron (MLP) learning model was generated, and trained using the features described above. It was desired that a representative of each major class of machine learning methods be used to evaluate the proposed approach. Therefore, a Bayesian network (BN), a support vector machine (SVM), and random forest (RF) classification models were all generated, one for each of the 4 word frequency schemes.

MLP was applied to predict whether or not the anonymous book belongs to the target author based on the one-class classification procedure. The MLP algorithm relies on a feed-forward ANN (artificial-neural-network) model which includes several layers of nodes and drives the input variables on to a proper output. The back-propagation technique is used for training the network, as presented in Algorithm 5.3. MLP has been applied using Weka version 3.7 and was trained using 'the backpropogation' procedure as shown in algorithms 5.1. It was performed as a classification module which takes advantage of the backpropagation algorithm. The backpropagation (as commonly used in machine learning community) process is an abbreviation for "backward propagation of errors" The flexibility of using MLP on Weka helps to enhance the understanding of the learning process as the network can be customised for a specific task by building the algorithm to suit the objectives. The network nodes all sigmoid because the ids of the classes are strings (author names). The MLP parameters were set based on initial empirical attempts; the number of epochs was set to 500, the rate of learning was 0.3, the momentum value was 0.2 (i.e. the default).

Algorithm 5.1 represents the steps of the back-propagation algorithm.

Back-propagation algorithm steps
1.Establish a feed-forward network with x inputs, k hidden units and y output
2. Intialise the learning rate r (0 or 1)
2. Assign weights to all nodes starting from small values randomly chosen
3. Repeat the following steps until the stop condition is acheived
(a) For each training example < x, atual > :
i.for each input x, compute the activation y= f(wx)
ii. For each unit compute the Er = y –actual
iii. Compute the bias b by the formula b = b + r * Er
iv. For all inputs I compute
Update the weight by w(i) = w(i) + Er * r * x(i)
where w = the weights vector, x= input vector, y= the correct output vale expected,
actual= the ouput of the unit and b = bias

The ability of this approach to detect ‘tampered-free’ which means the same author’s books. While the books of negative group assumed as tampered because it has not consistent with the model that was built.

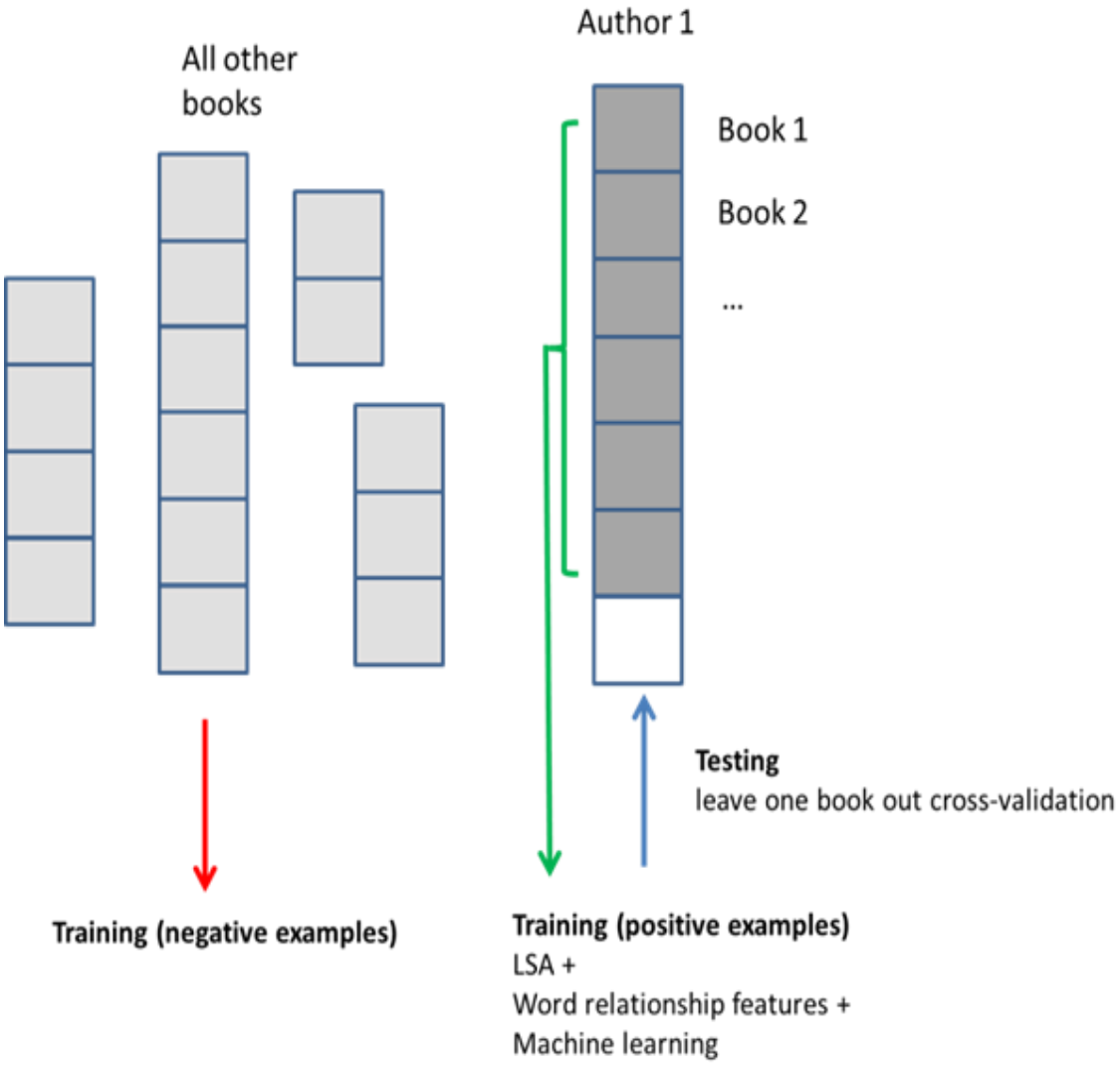

Figure 5.4. Positive examples, for each training set, consist of books for the particular author, while negative examples consist of all works not belonging to the author. (Source: The author)

Figure 5.4 represents the method of using positive and negative examples, the positives consist of books for the particular author. While negative examples consist of all works not belonging to the target author. The model of authorship of the intrinsic method for detecting plagiarism was generated based on LSA, Stylometry and MLP algorithms which were integrated with innovative features composition. An iterative learning and testing procedure was applied using the leave-one-out-cross-validation (LOOCV) technique to develop a robust

authorship detection model as depicted in figure 5.5. The procedure of (LOOCV) applied for each book in the target author dataset to train the model based on the proposed methods that described above. Figure 5.5 describes (using positive examples for the number of books in the target class and negative examples for the number of authors) the mechanism that was applied in order to perform on-class classification procedure.

Figure 5.5. Cross section for leave-book-out-cross-validation method (source: *http://robjhyndman.com/hyndsight/crossvalidation*)

5.2.2 Evaluation of the Model Performance

This approach was based on the method suggested by Zheng et al (2006) and modified to suite the requirement of this research proposed approach. Different metrics were proposed to evaluate the performance of the proposed approach. They include accuracy (ACC), sensitivity and specificity metrics. The metric of validity likelihood-ratio was calculated for each author. To explore how the performance metrics are reliable and accurate, the confidence interval for the prediction performance value for each book was presented in chapter 6.

5.3 Summary

This approach was proposed to identify if a specific text is written by the target author. This approach is based on deriving sets of features from MCW frequencies to measure the variation in the target author writing style. The core component of this approach is using the baseline frequencies of MCW and deriving different sets of features that reflect the in-depth distribution of MCW in each class. The frequencies of a particular MCW; the relative frequencies of all MCWs; the in-series proportional frequencies (e.g. 2nd / 1st, etc.); and the z-scores were calculated. The third and fourth sets of features mainly rely on estimating the probability from adjacent words. This kind of estimation is used to disclose connections between MCWs and expose their usage patterns. A Multilayer perceptron and three different machine learning techniquws were employed to generate the authorship prediction model. Therefore, a Bayesian network (BN), a support vector machine (SVM), and a random forest (RF) were all generated, one for each of the 4 word frequency schemes.

With regard to the application of the proposed techniques, one-class classification trained one author documents as positive examples and all other examples are assumed to be negative (outliers). Leave one-out-cross-validation (LOOCV) is used in order to switch the roles between testing and training data for full insights prediction performance evaluation. The algorithm of LOOCV is described in chapter 4.

LSA was limited to shrink the space and capture most common words usage patterns (no semantics properties are considered).

The next chapter discusses the performance metrics that were proposed to evaluate both extrinsic and intrinsic methods for plagiarism detection. Several metrics are discussed based on their contribution to the presentation of the experiment results. The next chapter will show the performance of the proposed methods.

Chapter 6: Results

The previous chapters (4 and 5) have presented the proposed extrinsic and intrinsic methods for plagiarism detection. Chapter 4 has proposed an extrinsic method for plagiarism detections targeting semantics. Chapter 5 has presented an intrinsic method of plagiarism detection to validate the text author. This chapter presents and discusses the obtained results for both methods.

Two different experimental designs were proposed and implemented due to the specificity of each method. The first experiment is named “extrinsic” because the comparison is performed with an external references collection which includes examples of documents from the same author. Training was performed for all authors, for experiment 1, there were 25 training sets and 292 test cases (all books). The processes of learning and testing were applied for 292 books based on their labelled classes. The second experiment is the named “intrinsic”, because training does not rely upon direct comparisons with external examples of documents of the author being queried. Rather, training generates a model summarizing the style of the author, which is based on examples of the author’s works, but which has a condensed metric for “style” that can act as a predictive model. This type of classification is termed as an on-class classification as it learns from a target class and all other classes are grouped and considered as negative examples.

The rest of the chapter is organised as follows. Section 5.2 presents some performance metrics related to machine learning approaches. Section 5.3 discusses the results of extrinsic method for detecting plagiarism. Section 5.4 discusses the results obtained by applying the proposed approach to intrinsic method for plagiarism detection.

6.1 Performance Metrics for Machine Learning

In the classification procedure, several metrics are used in order to assess the performance of classifiers. Thus, it is necessary to define some important elements with regard to the classifiers performance evaluation. There are common elements that are used in many evaluation formulas; true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These elements can be defined as follows.True positive rate (TP): number of predicted positive examples that are, in fact, positive

False positive rate (FP): number of predictive positive examples that are, in fact, negative

True negative rate (TN): number of predicted negative examples that are, in fact, negative

False negative rate (FN): number of predicted negative examples that are, in fact, positive

Sensitivity

There is a need to have a quantitative mean to assess the classification model, by evaluating its performance in respect of the assignment of the correct class value to the test books. The Sensitivity metric can be defined as the proportion of positive examples correctly classified as positives by the classifier out of all positive examples in the dataset. Formally it can be calculated using formula 2, as the number of true positives (TP) divided by the sum of true positives (TP) and false negatives (FN).

$$Sensitivity = TP/(TP+FN) \quad 6.1$$

Specificity

Specificity is the proportion of the true negative samples that were correctly detected by the detection system using formula 3. It suggests how good the detection method is at detecting the plagiarism free books (TN).

$$Specificity = TN/(TN+FP) \quad 6.2$$

Two more metrics such as precision or recall can also be used in evaluating the detection performance and commonly, both. Precision can be thought of as how true the predictions are, while recall can be thought of as "coverage" or how well the classifier reaches to all true instances. More formally using the same above elements,

$$Precision = TP/(TP + FP) \quad 6.3$$

$$Recall = TP/(TP + FN) \quad 6.4$$

The overall measure is name as accuracy and can be calculated based on formula 6.5

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad 6.5$$

6.2 Results of the Extrinsic Method for Plagiarism Detection

A series of experiments was performed to evaluate the efficiency of the proposed extrinsic method for plagiarism detection. . The proposed approach has evaluated a new use of LSA that aimed to enhance semantic plagiarism detection. Also to identify text authorship by capturing the usage patterns of MCW for each author as shown in figure 6.1. The figure depicted the variation in usage of the MCW "the" between four authors; Marion, Gissing, Parker and Louis.

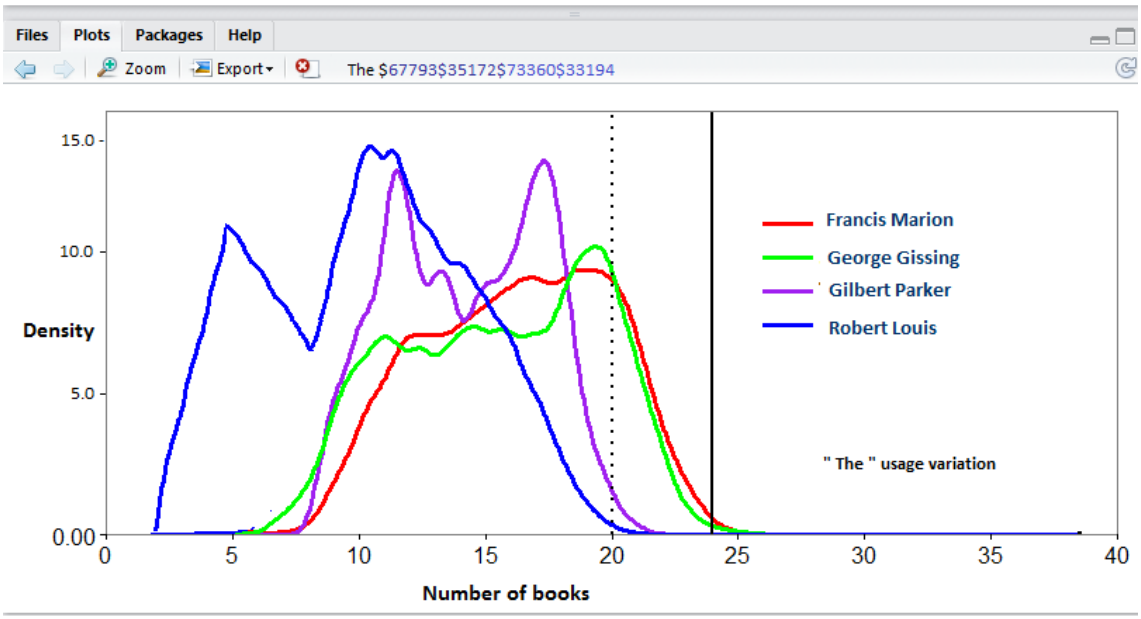

Figure 6.1. Presents the variation usage pattern of “the” between 4 authors

The new application of LSA in this approach is the use of most common words (words with high frequency) as additional features set with content words CW. The using of MCW has added a layer of stylometric analysis to capture the author writing style. LSA has incorporated in this approach to offer a deep linguistic analysis method that works on uncovering the latent association between terms. The combination of LSA and stylometry helps to build semantic models and captures the relevant patterns of MCW usage for text authorship detection. The method of boosting the MCWs weights assumed to be as a proactive discriminative step. It aims to discriminate between authors MCWs usage patterns in order to improve classifier performance as clarified in chapter 4.

6.2.1 The Results Presentation

The results of the proposed approach are presented in this section include; the results of the traditional LSA application as the MCW are neglected. Then the results of LSA based stylometry, where the weights of MCW are boosted. Both were incorporated with LSA based stylometry to show how the classification

algorithm influences the detection performance as shown in table 6.1. It was obvious that LSA based stylometry and SVM algorithm has outperformed the traditional LSA.

The resulting performance of the classifier of both methods is shown in table 6.1. The accuracy; the measure of overall performance is based on the two widely used metrics in classification problems: precision P and recall R as shown in formula 6.5. Precision (formula 6.1) is a metric to calculate the number of positive predictions divided by the total number of predictions for a class. Recall (formula 6.2) is a metric to calculate the number of positive predictions divided by the number of positive class values in the test data. In table 1 column one presents the names of the authors, column two presents the results from applying traditional LSA with a layer of machine learning. The application of traditional LSA has given the first conception of data behaviour and determined the directions for further investigation. The LSA was fine-tuned to include the most common words MCW as a group of high value features and the results were presented in column three.

Table 6.1. Presents the overall accuracy prediction results considering each class dataset based on traditional LSA and LSA based stylometry and SVM

Author/code	LSA+SVM	LSA based stylometry + SVM

Author/code	LSA+SVM	LSA based stylometry + SVM
Francis Marion	0.77	0.923
Robert Barr	0.75	0.891
Henry Seton	0.73	0.884
George Augus	0.78	0.944
Ralph Connor	0.79	0.912
Emerson Hough	0.76	0.891
Jerome Kapla	0.79	0.894
Frances Burnett	0.75	0.913
George Gissing	0.79	0.891
Kate Douglas	0.8	0.933
Edith Nesbit	0.81	0.962
Henry Rider	0.79	0.941
Hall Caine	0.73	0.86
Gilbert Parker	0.79	0.92
Lyman Frank	0.81	0.953
Arthur Conan	0.78	0.944
Andy Adams	0.79	0.91
Edith Wharton	0.69	0.882
Stanley John	0.79	0.921
Gertrude Atherton	0.76	0.927
Irving Bacheller	0.71	0.921
Grant Allen	0.76	0.938
Marie Corelli	0.74	0.919
Humphrey Ward	0.79	0.901
Robert Louis	0.77	0.891
Overall	**0.7688**	**0.9132**

The analysis was applied for each book from 292 books and validated using the leave-one-out-cross-validation method as described in algorithm 4.3. For each class (author), training used n-1 of the data set (Books) where n represents all the books in the author's dataset. Then the unseen element of the data set was used to test the method. Thus, for each book, the proposed approach was

trained using the 291 other books. It was clear that LSA based stylometry outperformed the traditional LSA which gives an answer for the research question that was formed for this part of research.

Two more classification algorithms assess the performance and calculate the results of the proposed method which uses the SVM-RBF (radial basis function) kernel, instead of the linear SVM. The kernel function aims to divert the objects into a higher dimensional space as a proactive step to apply the separation of classes in non-linear space. Also the SMO (Sequential Minimal Optimization) algorithm was implemented as a learning algorithm for SVM. SVM-SMO was used as a keen learner to grasp as much discriminative patterns as possible; it was defined as an eager learning algorithm (Luyckx and Daelemans, 2008). In order to obtain a best response from the SMO algorithm, several parameters need to be determined based on practical attempts. As all the features turned into their frequencies which meant all features were numeric and all checks processes were turned off. The complexity constant c was set to 0.98, the weight adjustments parameter was fitted to SVM outputs and the kernel for this method is polynomial. Table 6.2 presents the overall prediction results of each author (class), this results were built on many internal prediction processes. In table 6.2, the second column was extracted from table 6.1 to show the best performance between the three classification algorithms that relied on LSA based stylometry. The fourth column which includes the results from SVM-SMO application, this algorithm has outperformed the other two classification procedures; SVM and SVM-BRF. The results also revealed as shown in table 6.2 that SVM-RBF has performed better than the traditional SVM. All evaluation classification algorithms have relied on the captured features that resulted from LSA application instead of the brute-force features. Such compressed features enhance the run time efficiency for algorithms in particular SMO and also improve the prediction accuracy.

Table 6.2. Presents the prediction accuracy of 25 authors (classes) that relied on SVM, SVM-BRF and SVM-SMO

Author/code	LSA based stylometry + SVM	LSA based Stylometry +SVM-RBF	LSA based Stylometry SVM-SMO
Francis Marion	0.92	0.928	0.932
Robert Barr	0.89	0.901	0.941
Henry Seton	0.88	0.899	0.947
George Augus	0.94	0.943	**0.961**
Ralph Connor	0.91	0.917	0.942
Emerson Hough	0.89	0.899	0.929
Jerome Kapla	0.89	0.902	0.929
Frances Burnett	0.91	0.9	0.918
George Gissing	0.89	0.89	0.937
Kate Douglas	0.93	0.937	0.93
Edith Nesbit	0.96	0.971	**0.969**
Henry Rider	0.94	0.948	**0.952**
Hall Caine	0.86	0.87	0.921
Gilbert Parker	0.92	0.924	0.931
Lyman Frank	0.95	0.951	**0.95**
Arthur Conan	0.94	0.973	**0.979**
Andy Adams	0.94	0.948	**0.956**
Edith Wharton	0.88	0.902	0.911
Stanley John	0.92	0.928	0.935
Gertrude Atherton	0.92	0.92	0.938
Irving Bacheller	0.92	0.93	0.936
Grant Allen	0.93	0.949	**0.955**
Marie Corelli	0.91	0.927	0.939
Humphrey Ward	0.9	0.912	0.942
Robert Louis	0.89	0.912	0.934
Overall	**0.9132**	0.92324	**0.94056**

6.2.2 Discussion

As explained in chapter 4 MCWs are always eliminated in the pre-processing process because MCWs have no contribution for the text meaning that is targeted by LSA. However MCWs from another point of view represent a salient

component in stylometric methods to identify authors writing style. As presented LSA based stylometry results outperformed the traditional LSA as shown in table 6.1. The prediction performance has obviously improved after the using of SVM with the fine-tuned LSA. The application of LSA based stylometry behaves on a monotonic trend of increasing predictive accuracy. This trend was not only a general trend, but in fact observable for every single author. Some authors such as those who were highlighted in red in table 6.1; George Augus, Edith Nesbit, Henry Rider, Lyman Frank and Arthur Conan have scored high prediction accuracies. These authors have scored 94.4%, 96.2%, 94.1%, 95.3% and 96.4% respectively which indicated that high number of their books are correctly predicted to their right labels. Other authors are scored less such as the novel titled The Shadow of a Crime (123,977) which was written by Hall Caine. This book is written in a different style by the same author to add a flavour of excitation.

There are substantial variances between studies in terms of how LSA is performed. This variability in dealing with LSA prevents researchers from using the parameters that are essential to LSA successful performance (Haley et al., 2007; Jorge-Botana et al., 2010). These parameters can include: the elimination of most common words (stop list), use of the weighting and the use of the value of k for dimensionality reduction applied by singular value decomposition SVD. Also in most authorship analysis and plagiarism detection there is no unified platform that can be used for evaluation. Zhao and Zobel (2005) and Mozgovoy, Kakkonen, and Cosma, (2010) stated that there are many systems that were proposed for authorship analysis or plagiarism detection relying on different classification approaches such as support vector machines or natural language processing techniques. These techniques do not have a standard evaluation platform as no consistent corpus is available. Most of existing techniques have relied on words n-grams or character n-grams to detect the similarity between two texts.

In order to put the obtained results into perspective, this work compares the results of LSA with what was reported by Ceska (2008) who has also used LSA

for plagiarism detection and dimensionality reduction. The research has also used n-grams to compare the LSA performance. Table 2.1 has presented the results that were obtained from traditional LSA and LSA based stylometry application. Some examples of successful BOW based plagiarism-detection methods have used advanced machine-learning techniques, such as support vector machines (SVM) (Zechner et al., 2009). LSA is actually a fairly untapped method in this regard, with very few cases of application to the task of detecting plagiarism discoverable in the literature (Ceska, 2008).

The stylometric method appears to be very successful at discriminating between authors. However, most of the scholarly work conducted so far has focused not on plagiarism detection, but on academic scrutiny of famous authors (Holmes, 1998; Juola, 2006). It is quite possible that the method has, in fact, significantly more to offer to the field of plagiarism detection.

Another observable trend is that the predictive accuracies for all authors' classes were more uniform, between authors. This was primarily due to the fact that, for authors with comparatively low predictive accuracy for novels using traditional LSA, the predictive accuracy when applying LSA based stylometry has increased.

6.3 **Results of the Intrinsic Method for Plagiarism Detection**

This section presents the evaluation results of the proposed intrinsic method for plagiarism detection which uses no direct comparison to a reference collection. Rather, training generates a model summarizing the style of the author, which is based on examples of the author's works, but which has a condensed metric for "style" that can act as a predictive model. This type of classification is termed as

an on-class classification as it learns from a target class and all other classes are grouped and considered as negative examples. The proposed approach is based on capturing the underlying patterns, by using stylometric analysis to generate characteristic features of an author's body of work. The captured patterns are used to identify if the characteristics of a written text has changed somehow. The approach has relied on four techniques BOW, LSA, stylometry and MLP. As stated in chapter 5 the MCWs are used to represent the books of 25 authors that comprised the corpus of English novels (CEN).

LSA was used in this approach for shrinking the features space and capturing the patterns of MCW usage. As discussed, because the literature dealing with Stylometry finds an unusual usefulness for function words, and because these function words are often the MCWs, the intrinsic model actually deals primarily with the frequencies of the MCWs found in the Corpus of English Novels (CEN), divided up by author. The raw or "baseline" frequencies comprised the first set of features (Fs1), these frequencies were calculated for each book, separately. Once these baseline frequencies were calculated, their frequencies were expressed as proportions of the total word count to form the second feature set, or Fs2, (e.g, frequency of each MCW divided by the Total frequencies of all MCWs). The third features set (Fs3) was based on calculating in-series proportional frequencies (2nd / 1st, etc.). Furthermore the z-value of the variance respective to the overall set of documents for the author, are all calculated, for each book as explained in chapter 5 has formed the features set Fs4.

Several measures of evaluation performance were used such as sensitivity, specificity and likelihood ratio. The likelihood ratio is an independent metric which works by putting more confidence on the results and weakens the error potential. This calculation can be performed by applying formula 6.6 to rule-in that the text is tampered, while the formula 6.7 was used to rule-out that the text is tampered. Another metric named the confidence interval metric was used to express the reliability and validity associated with a proposed sampling method. The confidence of the classifier prediction performance can be defined as an

indicator of the reliability of the detection results (Morrison, 2010). In other words, the confidence interval represents how precise and stable are the performance measurements when the experiments are repeated. The confidence interval metric was used at the individual book level in order to assess the performance of the proposed approach on each author set. On the other hand the negative LR^- (likelihood ratio) as shown in formula 6.7 was used to assess the performance of the proposed approach on all authors' datasets. Both metrics enhance the credibility of the method on an individual book basis as well as entire author's dataset.

$$\text{Likelihood Ratio (LR)}^+ = \text{Sensitivity} / (1\text{-Specificity}) \quad 6.6$$

$$\text{Likelihood Ratio (LR)}^- = (1\text{- Sensitivity}) / \text{Specificity} \quad 6.7$$

Table 6.3. The standard confusion matrix

	Positive Class	Negative Classes
Classified as Positive	True Positives TP	False Positives FP
Classified as Negative	False Negatives FN	True Negatives TN

The confusion matrix in table 1 describes the process of classification model (or "classifier") on a set of test data for which the classes are identified. The primary parameter adjusted across a range of values in order to explore predictive capability was the MCWs. Therefore, frequencies of the top 10, 20, 50, and 100 MCWs were calculated.

6.3.1 The Results Presentation

Tables from 6.4 to 6.7 present the prediction performance of four corpora (authors) as sample indicators of the prediction performance of the proposed

approach. The results shown are for the complete analysis included the top 10, 20, 50, and 100 most common words.

Table 6.4. The prediction results on the "Gertrude Atherton" dataset

Book	Prediction correct	Confidence (0-1)
1906 rezanov	+	**0.968**
1900 senator north	+	0.936
1921 the sisters-in-law	+	0.948
1922 sleeping fires	+	**0.98**
1888 what dreams come	+	**0.95**
1902 the splindid idle forties	+	**0.969**
1918 the white morning	+	**0.978**
1898 the valiant runaways	+	**0.98**
1900 the doomswoman	+	0.945
1919 the avalanche	-	0.884

In Tables 6.4 to 6.7, a "+" sign in the "Prediction correct" column indicates a correct prediction of the author, while the confidence of the prediction made (whether correct or not; for incorrect predictions this is the confidence held that the incorrect prediction was in fact correct) is indicated in the rightmost column. The confidence interval shows the level of credibility on the prediction results and used to infer that the true value lies between determined two points. Most studies rely on the 95% confidence interval which interpreted to be occurred between the values (0-1) (Field, 2013).

The tables of the prediction results presented books that scored higher, smaller or equal to 0.95 concludes that all these values belongs to the interval (0-1). If repeated samples were taken and the 95% confidence interval was computed for each sample, then the performance of the proposed approach can be described as 95% generalised to real-world samples. The confidence-level values express that the prediction ability of the intrinsic plagiarism proposed

approach is reliable and the proposed approach is likely to get good performance on other samples.

Table 6.5. The prediction results on the Henry Seton corpus

Book	Prediction correct	Confidence (0-1)
1897 in kedar's tents	+	**0.955**
1900 the isle of unrest	+	0.942
1892 the slave of the lamp	-	0.887
1894 with edged tools	+	**0.953**
1902 the vultures	+	0.948
1892 from one generation to another	+	**0.968**
1901 the velvet glove	+	**0.956**
1895 the sowers	+	**0.987**
1913 roden's corner	+	0.946
1904 the last hope	+	**0.983**
1903 barlasch of the guard	+	**0.968**
1895 the grey lady	+	**0.982**

Table 6.6. The prediction results on the Lyman Frank corpus

Book	Prediction correct	Confidence (0-1)
1908 dorothy and the wizard in oz	+	**0.972**
1901 american fairy tales	+	**0.966**
1906 aunt jane's nieces	+	**0.954**
1916 mary louise	+	0.886
1911 aunt jane's nieces and uncle john	+	0.893
1900 the wonderful wizard of oz	+	**0.972**

1910 the emerald city of oz	+	**0.991**
1912 sky island	+	**0.986**
1902 the surprising adventures	+	**0.973**
1915 the scarecrow of oz	+	**0.979**
1907 ozma of oz	+	**0.981**
1906 aunt jane's nieces abroad	+	**0.984**
1912 aunt jane's nieces on vacation	+	**0.978**
1903 the enchanted island of yew	+	**0.981**

The above results were obtained from the proposed approach which was performed on a per-book basis, based on the proposed statistical features sets. Two sets of features; Fs3 and Fs4 were mainly based on estimating probabilities from adjacent words. This kind of estimation has played a strong role in disclosing important connections between MCWs and exposing their usage patterns. As stated the use of MCWs is a frequent practice for machine-learning authorship models. The use of statistical properties of words is related to these models, but in this case is distinct from multivariate approaches which focus on Stylometry. The proportions of each other, in sequential order (e.g., the #2 MCW's frequency proportion was divided by the #1 mcw's frequency-proportion) which forms the third feature set Fs3. This is considered as an important estimator for adjacent words' connection; this feature was named "in-series frequency ratios" and it constitutes one of the novel contributions of this research.

Table 6.7. The prediction results on the Humphrey Ward corpus

Book	Prediction correct	Confidence (0-1)
1884 Miss Bretherton	+	0.978
1900 Eleanor	+	0.969
1898 Helbeck of Bannisdale 2	+	0.98
1913 The Mating of Lydia	+	0.996
1916 Lady Connie	+	0.984
1911 the case of Richard Meynell	+	0.876

1906 Fenwick's Career	+	0.976
1888 Robert Elsmere	+	0.992
1908 The testing of Diana Mallory	+	0.981
1896 Sir George Tressady 2	+	0.996
1913 The Coryston family	+	0.979
1915 a great success	+	0.981
1914 Delia Blanchflower	+	0.993
1905 The marriage of William Ashe	+	0.984
1894 Marcella	+	0.985
1881 Milly and Olly	+	0.979
1903 Lady Rose's Daughter	-	0.997

Table 6.8 has presented the overall performance of the intrinsic method for plagiarism detection. Three types of metrics were presented including sensitivity, specificity and negative likelihood ratio (LR^-). The advantage of using the negative LR^- metric was to show the stability of the proposed model. This negative type of LR has been used to rule-out plagiarism. The metric showed the stability of the outcomes and reflected the reality of the results of the authorship cases. The interpretation of likelihood ratios values was based on balancing the sensitivity and specificity values.

Table 6.8. The overall results for all 25 classes using the intrinsic plagiarism detection proposed approach

Code	Author Name (class)	Sensitivity	Specificity	LR^-
A	Henry_Rider	0.96	0.998	0.0401
B	Kate_Douglas	0.928	1	0.0720
C	Hall_Caine	0.333	1	0.6670
D	Edith_Nesbit	1	1	0
E	Irving_Bacheller	1	1	0
F	Lyman_Frank	1	0.992	0
G	Marie_Corelli	1	0.996	0

H	Gilbert_Parker	0.941	1	0.0590
I	Henry_Seton	0.916	1	0.0840
J	Ralph_Connor	1	1	0
K	Humphrey_Ward	0.941	0.996	0.0592
L	Edith_Wharton	0.909	0.992	0.0917
M	Emerson_Hough	0.777	0.992	0.2247
N	Grant_Allen	1	1	0
O	Robert_Barr	1	1	0
P	Jerome_Kapla	0.9	0.996	0.1004
Q	Frances_Burnett	1	1	0
R	Andy_Adams	0.8	1	0.2
S	George_Augustus	0.8	0.996	0.2008
T	Stanley_John	1	1	0
U	Gertrude_Atherton	0.9	0.996	0.1004
V	Robert_Louis	0.888	1	0.1120
W	George_Gissing	1	1	0
X	Arthur_Conan	1	0.992	0
Y	Francis_Marion	0.923	0.996	0.07731
Overall results				**0.08355**

The larger the positive likelihood values the greater the indication that the text was tampered. While the smaller the likelihood value, the greater the indication that the text was tampered-free. As stated before, a high value of specificity was more important than sensitivity in plagiarism and authorship detection, so a high specificity value indicated that the text was tampered-free (it always occurs with high values of TN). This was compatible with using the LR metric, a smaller likelihood value indicated that the text was tampered-free, which means a high value for TN.

Table 6.9. Presents the misclassification error (Miss-E) for each set of the proposed features based on four classification algorithms

	SVM				MLP			
F-type	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
Miss-E	0.281	0.221	0.191	0.213	0.063	0.061	0.003	0.006
	BN				RF			
F-type	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
Miss-E	0.137	0.125	0.064	0.075	0.322	0.297	0.178	0.193

The most informative features for the neural network based model (multilayer perceptron or MLP) were those dealing with standard deviation. Especially, the features set of in-series frequency ratios of MCWs (F3 as shown in Table 6.10, indicated in red). This represents one of the most original contributions of this work, as the relative frequencies of words (as opposed to the raw frequencies) have not yet been reported in the literature.

To investigate different features' sets and classification algorithms, several authorship verification tasks were proposed. As stated above four feature sets were constructed, F1 represents the frequencies, F2 expressed the frequencies as proportions of the total word count.

F3 represents in-series proportional frequencies (2nd / 1st, etc.) and finally the z-value of the variance respective to the overall set of documents for the author. Two types of analysis procedures was applied, firstly each features set was examined separately the mis-classification error based on four classification algorithms was calculated. Table 6.9 presents the performance of each classifier on each proposed set of features. The third set of features has scored the lowest mis-classification error (MSE) value for all classifiers algorithms. This set of features presents one of the original contributions for this research. Secondly, the first featureset Fs1 was examined then the second feature set Fs2 added to Fs1 to form the second feature set (Fs1+Fs2). Fs3 set was added to form the third feature set (Fs1+Fs2+Fs3). The fourth feature set contains all four types of features (Fs1+Fs2+Fs3+Fs4). This incremental method was

chosen because it represents the evolutionary sequence of style features that measures the text density (Zheng et al, 2006) as shown in table 6.10.

Table 6.10. The performance of four ML methods based on different sets of feature compared to other classification algorithms based on detection accuracy

	SVM	MLP	BN	RF
FS1	0.7815	**0.8823**	0.8611	0.8107
FS1+FS2	0.8021	**0.897**	0.8746	0.8557
FS1+FS2+FS3	0.8651	**0.9396**	0.9198	0.8772
FS1+FS2+FS3+FS4	**0.8885**	**0.9715**	**0. 9257**	**0.8625**

The two application procedures (features were analysed separately and in accumulation way) added new types of features to the existing sets. Four classifies were generated as classification algorithms including support vector machines (SVM), multilayer perceptron (MLP), Bayes network (BN) and random forest (RF) respectively. Leave-one-out-cross-validation was used to estimate the accuracy of the classification model. It is obvious from table 6.10 that multilayer perceptron (MLP) has outperformed the other algorithms by scoring (0.97) as a prediction accuracy value using all groups of features. The experimental design investigates the impact of analysing different numbers of MCWs on the performance of authorship verification. Four tasks have been set up to use 10, 20, 50, and 100 MCWs, respectively. In each turn, the classification algorithm was trained to generate the model; this procedure was iterated for each classification algorithm and for each author dataset.

In order to evaluate the performance of the intrinsic detection method, four metrics including specificity, sensitivity, accuracy and calculating the mis-classification error (MSE) for each classifier as shown in table 6.11. These metrics are assumed to be the main evaluation metrics that are used in authorship analysis, data mining and plagiarism detection (Zheng et al., 2006).

The confusion matrices giving the overall prediction performance of the proposed method were presented in table 6.11 as well.

Table 6.11; concludes the performance of four classifiers based on four metric techniques which included MSE, overall accuracy, specificity and sensitivity. MLP outperformed all other algorithm using the four sets of features separately as well as all of them together, which confirmed the efficiency of employing jointly Stylometry and MLP. FS3 features set was the dominant set that affected the performance of all machine learning algorithms, in particular the MLP, as shown in Table 11 (highlighted in red), which presents the effects of each features set on the four machine learning algorithms that were used, by calculating the MSE for each one.

Table 6.11. The averaged results for four classification algorithms: MLP, BN, SVM and RF

	Multilayer Perceptron				Bayesian Networks				Support Vector Machine				Random Forest			
Misclassification error	0.0240				0.0514				0.1268				0.3802			
Accuracy	0.9715				0. 9257				0.8885				0.8625			
Specificity	0.9090				0.6818				0.0.608				0.3703			
Sensitivity	0.9857				0.9703				0.9294				0.7663			
Confusion matrix for the experiments Labels 1: tampered 2: tampered-free	True Labels	Estimated		Labels Totals	True Labels	Estimated		Labels Totals	True Labels	Estimated		Labels Totals	True Labels	Estimated		Labels Totals
		1	2			1	2			1	2			1	2	
	1	277	1	1	1	262	7	269	1	224	20	244	1	141	68	209
	2	4	10	2	2	8	15	23	2	17	31	48	2	43	40	83
	Totals	281	11	292	Totals	270	22	292	Totals	241	51	292	Totals	241	108	292

6.3.2 Discussion

The proposed approach has used 4 types of feature sets as described above and generated a number of different machine learning models, which were trained for evaluation purpose. A representative of each major class of machine learning methods was used, therefore, a MLP, a Bayesian network (BN), a support vector machine (SVM), and a random forest (RF) were all generated, one for each of the 4 word frequency schemes.

In addition to the main metric to measure the approach prediction performance, three more statistics are involved in evaluating the effectiveness of the proposed including sensitivity, specificity and the likelihood ratio. The likelihood ratio is a valuable and concrete method of evaluating the stability of the proposed models. Two main factors have driven the effectiveness of successful detection approaches so far which are the features types used and the machine learning algorithms that were employed to develop the models. When the features were examined separately the third feature set (Fs3) was outperformed by the other feature sets as shown in table 6.9. When the accumulation mechanism was applied for the four classifier algorithms used, the accuracy increases as more features are added, in a proportional relationship as shown in Table 6.10. The highest value of accuracy was obtained from MLP based on LSA and Stylometry methods which scored (0.97). The dominance of the MLP approach was enhanced when all features sets were added in particular the sets of features that were based on the probability of adjacent words (F3 and F4). The results have given strong evidence that the efficiency of the methods that are predicted from the word connections as can be seen clearly in table 6.10.

MLP has outperformed other algorithms in terms of MSE, Sensitivity, specificity and accuracy as clarified in table 6.11. On the other hand, it is also not entirely obvious why this should be the case. Certainly, neural networks have a long and important history in the field of stylometry. They were among the earliest machine-learning models that were applied to the tasks of encapsulating author styles and predicting authorship (Matthews and Merriam, 1993; Tweedie, Singh

and Holmes, 1996). The second most effective machine-learning method, given the features generated, was a Bayesian network. Bayesian networks actually have the oldest roots of all machine-learning methods, for the task of authorship prediction. One of the best-known and longest-standing of the multivariate approaches for authorship analysis was based on a rigorous Bayesian framework linking common functional words, as in the famous study of Mosteller and Wallace (1964). This early work did not constitute machine learning; however, later works such as Genkin & Lewis (2007) employed Bayesian regression. In the proposed approach, the Bayesian networks took only a fraction of the time to generate the model that was taken by the Multilayer Perceptron. Although generally the compilation time of a machine learning model was not all that important, provided that it can be done in a matter of hours or days, the difference in time was at least one order of magnitude. Also, surprisingly, SVMnd Random Forests performed worse than the MLP. As stated in section 6.2.2 there is no standard or unified evaluation platform for system evaluation in authorship analysis or plagiarism detection. This is due to several factors such as the lack of consistent corpora and the variation in using different methods which makes the comparison of the obtained results with others difficult. However, to put the proposed work in a track for future evaluation, the intrinsic proposed method has followed a work that was conducted by (Zheng et al., 2006) to investigate several sets of features and algorithms. They first created four feature sets contained structural, syntactic, lexical and content-specific types of features. In their study they accessed publicly available newsgroup messages that were selected and collected for the test bed in this study, to detect the writing styles. The proposed method is based on the method suggested by Zheng et al (2006).

6.4 **Conclusion**

This chapter has presented the results of the proposed intrinsic and extrinsic methods the the detection of plagiarism. In order to evaluate the efficiency of these methods, a series of experiments were conducted on 25 authors' datasets belongs to the corpus of English novels (CEN). The experimental

results determined that the proposed approaches were able to detect the author's class for extrinsic plagiarism detection (when an external reference collection is available for comparison), and intrinsic plagiarism detection (when no external references collection is available). SVM-SMO has outperformed traditional SVM and SVM-RBF using the extrinsic method, where MLP outperformed Bayes network (BN), SVM and random forest (RF) using the intrinsic method. Several parameter settings were applied which have impacted the performance positively. The most informative features for the neural network machine learning model (MLP) were those dealing with standard deviation, especially standard deviation of the in-series proportions of MCWs. This represents one of the most original contributions of this work, as the relative frequencies of words (as opposed to the raw frequencies) have not yet been reported in the literature.

The reasonable performance that was achieved for all authors' classes in both approaches is encouraging for the future use of these methods in a real-world context. The results have evidenced the ability of the proposed approaches in tracing the identities of authors for many applications, especially in academic bodies and publication firms.

The next chapter discusses the main conclusion of the thesis including the summary, contributions and limitations of the research. It also discusses the future work.

Chapter 7: Conclusion

7.1 Introduction

The previous chapter discussed the results of this research including the evaluation of the performance of the two proposed approaches. Several metrics were discussed based on their role in presenting the experiments results. . The chapter has shown the performance of the proposed approaches and detailed discussion for the presented results is produced.

The extrinsic method for plagiarism detection has achieved a high score in capturing the patterns of semantics and identifying the authorship characterisation with the availability of a reference collection for comparison. The fine-tuned procedure of LSA application has enhanced the performance of the proposed method as shown in chapter 6.

Using the Intrinsic method for plagiarism detection several machine learning models were generated and different evaluation metrics were used to measure the performance of the proposed method. The results revealed the proposed intrinsic method had scored highly when using one of the proposed features sets which were proposed for this research. The most informative feature set were those dealing with standard deviation, especially standard deviation of the in-series proportions of most common words. This represents one of the most original contributions of this work as clarified in chapter 6.

The rest of the chapter is organised as follows: Section 2 provides a summary of the research. Section 3 discusses the contribution of the research in relation to the research objectives. Section 4 discusses the research limitations. Section 5 discusses the future work.

7.2 **The Research Summary**

The motivations behind this research are related to the lack of a precise practical framework to address the limitations of existing plagiarism detection methods in revealing the semantic relationships between texts and identifying the text source. The interaction process between users and detection tools do not go much further than highlighting the similarity between submitted texts and the repositories of plagiarism tools. In fact the tools have strayed far from their mission to protect the scientific environment and emphasise ethical concepts. On the contrary they are forcing the users to find different ways in order to deceive the checking algorithms in some cases. Furthermore they have failed to influence the awareness of users regarding plagiarism as they are targeting the imitation of the language rather than the meaning of the text. In the research environment, researchers are expected to develop novel knowledge in a particular discipline; therefore the usage of words is a way of expressing the thoughts, innovations, suggestions, approaches and outcomes associated with that knowledge. The existing plagiarism detection tools targeted the patterns of language without considering the core of the novel knowledge or how this knowledge was developed. Yilmaz, (2007) argued that the issue of "borrowing good English" in order to describe novel ideas and knowledge should be an acceptable practice in some sectors. However the essential concern is whether the current plagiarism detection tools can differentiate between two texts that used the same words but where the knowledge content is different. The limitations in existing detection tools were evident even in the detection of the words themselves. They are limited when dealing with context that can be embedded in a word. Synonyms (i.e. multiple words for the same meaning) are a key concern for plagiarism detection methods. Most of the current techniques can easily be tricked by experienced writers through text obfuscation and manipulation (Afroz et al., 2014).

For addressing the limitations of extrinsic detection methods for plagiarism, this research has incorporated bag of words (BOW), latent semantic analysis for semantics detection with stylometry and support vector machines (SVM)

techniques. The proposed method was named extrinsic because training includes examples of documents from the same author as the document being classified. The BOW technique represents the first step for generating the initial features set which produced high dimensionality in vector space. Stylometry is a profound component for the proposed approaches; the primary stylometric parameter was the frequencies of most common words (MCWs). This parameter was used across a range of values, in order to explore the predictive capability for extrinsic and intrinsic methods. With regard to the extrinsic validation experiment, the work in this thesis focused on developing a supervised learning-based plagiarism detection concept that maximises the semantics and stylometric detection performance while minimising false positives (i.e. where the cases are detected as plagiarised but in fact they are not).

In order to efficiently address the problem of semantics detection, latent semantic analysis LSA was shown to be the most suitable method for plagiarism detection approaches owing to the following reasons:

1. The method can identify the text meaning by revealing the latent associations between words in order to capture the semantics patterns.

2. It is considered to be a means of shrinking the feature space for a BOW. By reducing the number of extracted dimensions to the barest minimum, noise is excluded and the amount of data and memory consumption is significantly reduced. The transitivity attribute of LSA facilitates capturing the specific co-occurrence of words.

As was evidenced by the literature and clarified in chapter 3 the traditional application of LSA always eliminates the most common words or commonly function words. The set of words is eliminated because they don't make any contribution to the meaning of the text. However such a class of words is considered to be a salient discriminative feature for characterising the document authorship. For this research, the LSA application was fine-tuned to take-in the stylometric features (most common words) in order to characterise the

document authorship as described in chapter 4. The results revealed that LSA based stylometry has outperformed the traditional LSA application. Support vector machine based algorithms were used to perform the classification procedure in order to predict which author had written the test book. The proposed approach answered the first research question successfully by addressing the limitations of semantic characteristics and identifying the document source by assigning the test book to the right author in most cases.

In order to address the limitations of the existing intrinsic methods for detecting plagiarism, the proposed method was built on a notion of extending the authorship verification methods. The method named intrinsic because training does not, in strict terms, rely upon direct comparisons with external examples of documents for the author being queried. Rather, training generates a model summarizing the style of the author. This model is based on examples of works of the author, but which has a condensed metric for “style” that can act as a predictive model. Stylometry as stated earlier is the linguistic root for the identification of text authorship by employing computational methods (Craig, 2004). The key component of this approach was to explore the effect using only a subset of the most common words (MCWs), which were subject to specific computational functions. This approach has relied on incorporating latent semantic analysis and multilayer perceptron. The idea of using a subset of MCWs was built on a notion that the writing style of authors can be patterned uniquely based on MCW usage. The disadvantage of intrinsic methods for plagiarism detection is the small size of the training corpus which can be as little as a handful of documents. A new application of Latent Semantic Analysis (LSA) has been proposed. The application of LSA addresses the problem of the number of documents and authors that ccurred with intrinsic methods for detecting plagiarism. The motive behind the use of latent semantic analysis is the ability to exploit word co-occurrence patterns which then allows a set of latent associations to be derived and also enables a shrunken vector space to be generated. Probably the most important aspect of machine-learning methods is the choice of features. A number of different machine learning models were generated, and trained using computational features. The machine-learning

method found to be the most effective at making predictions, was the artificial neural network (multilayer perceptron). Using the frequencies of MCWs as a baseline parameter for this approach drives the process for devising three more feature set derivatives which were used both separately and in an accumulation mechanism. The standard deviation of the in-series proportions of MCW feature sets has scored the best results among other feature sets. This represents one of the most original contributions of this work, as the relative frequencies of words (as opposed to the raw frequencies) have not yet been reported in the literature. Based on the obtained results the the second research question has been answered as the proposed method was able to predict if the test book was written by the target author or not.

7.3 Contribution to the Knowledge

This thesis has made a number of novel contributions as briefly described in chapter 1. They were:

The thesis has introduced a novel approach based on the integration of a number of well-known techniques in order to address the issues relating to plagiarism detection. More precisely the issue of text identification with reference and without references is identified and addressed by proposing two different but complementary methods. The contribution to the knowledge of this research can be described based on the achievements for each method.

- Considering the extrinsic method for plagiarism detection:
    a. LSA was fine-tuned to take-in MCW features set as a main component for LSA applications. The MCWs are commonly eliminated before any LSA application during the pre-processing stage due to their high frequencies. Even if they retained for some applications, they will be given trivial weight. The fine-tuned procedure was based on a notion that MCWs that appear in one class over others can play a significant discrimination role. With this method the MCWs were retained and the weighting method was fine-tuned to reflect their importance in each class(author's

dataset) as clarified in chapter 4. Classification algorithms can be defined as methods that can perform “only as well as the data put in”. Preparing a well-formed feature set can substantially enhance the classification accuracy. An SVM using three different classification methods was used to select the best model based on their classification accuracy.

This approach has addressed the limitations of current approaches in two aspects:

- Improving the capability of capturing the text semantics
- Improving the capability to identify the text source by capturing the authorial attributes patterns.

b. The complementary combination of BOW, LSA and stylometry also represents a new approach to address the limitations of extrinsic methods for plagiarism detection. The complementary scenario can be briefly described by clarifying the roles of each component (technique) and how the integration process enhances its performance. BOW represents the first step for generating the initial features set which produced high dimensional vector space. LSA works on analysing the text contents without considering the stylistic attributes for author identity. This works in a contrary manner to stylometry which identifies the text identity without considering the text contents. The proposed mediation process that was applied, made the integration between these two well used techniques for analysing the text in a different way reasonable.

   - LSA was used for dimensionality reduction by applying its internal algorithm SVD to reduce the high dimensionality of the BOW technique. It also performs a deep text analysis to reveal latent associations between words to capture text semantics. In spite of LSA’s ability to capture the semantics of the text, it was not able to capture the

authorial attributes. This weakness has reduced ts classification accuracy.

- In contrast stylometry performs superficial analysis that quantifies the statistical attributes for text authorship applications. Stylometry in the proposed method has enhanced the performance of LSA for classification tasks by engaging MCW features.

- Another part of this research contribution is developing an intrinsic method for plagiarism detection. Intrinsic methods for plagiarism detection use no direct comparison to an external document collection. Instead, it uses stylometric analysis to generate characteristic features of an author's (i.e. a suspected plagiariser's) body of work. The new writing output from an author is compared to old examples for that author in order to capture variation in the writing style. This method has worked for verifying if the target author wrote the suspicious document. The method based on proposing a new technique for deriving discriminating sets of stylometric features which is called features engineering (FE). The comparison is applied without relying on a reference collection. This approach has relied on using a set of MCWs to represent each author's dataset.
    a. LSA was used for the first time for this task to capture the MCWs distribution patterns and optimise the retention of related MCWs. The optimal features set that resulted from LSA was used as a baseline to derive more feature sets.
    b. Sets of new statistical features were derived based on the MCW frequencies used as inputs for a multilayer perceptron algorithm. One of the derived feature sets outperformed others, the feature was named the "in-series frequency ratios" of MCWs. This represents one of the most original contributions of this work, as the relative frequencies of words (as opposed to the raw

frequencies) have not yet been reported in the literature. This is clarified in chapter 6.

- The research proposed a novel experimental methodology for testing the performance of extrinsic and intrinsic plagiarism detection methods. The experiments covered thousands of machine learning processes based on leave-one-out-cross-validation (LOOCV).
- To the author's knowledge, this research is the first attempt to address the limitations of extrinsic and intrinsic plagiarism detection methods in the same study.

7.4 Research Limitations

In any research, the presented results need to be considered within the context of the limitations. In addition, the procedure of creating and answering specific research questions usually produces more concerns that need to be explored through further research. This research proposed methods were limited to the following:

It is important to note, first and foremost that existing plagiarism corpora very rarely consist of actual plagiarism cases. Naturally occurring plagiarism cases are difficult to come by. Acquiring real plagiarism cases is often imbued with legal, social and ethical issues in addition to other technical concerns. The legal and ethical concerns of using real plagiarism cases in a corpus is in fact an obstacle as it requires the consent of both the original author and the plagiarist. It is needless to say that it is rare for potential plagiarists to actually admit to a plagiarism offence. The technical concerns of using real plagiarism cases relate to the difficulty in getting large enough corpora for more extensive research. The aforementioned concerns have resulted in making the task of defining an appropriate plagiarism detection experimental corpus a challenge. The development of such experimental collections that are sophisticated enough to model the research problems, to serve as a standard corpus is still an obstacle in this field (Afroz, 2013). An alternative to offset the lack of particular corpora is

to use one of the available collections which are related to text analysis. Such corpora are freely available for research purposes subject to some requirements. The problem for every researcher when using such corpora (to different extents) is the lack of complete real-world depiction. For example, the corpus of English novels which is used to track the language changes in written text and was used for the purpose of this research provide 292 unlabelled novels. Processing steps were applied for the corpus of English novels to facilitate its usage for our research experiments as each author was labelled as a separate dataset.

7.4 Future Work

A number of future research directions are identified as follows:

Extrinsic plagiarism detection

For future work on extrinsic methods for plagiarism detection the author intends to apply another form of machine learning algorithms called Meta Learning (Unmasking method) for each author dataset. The Unmasking method helps in reflecting the consistency of the applied classification model. The method of Unmasking is an approach proposed by Koppel et al., (2007). This method relied on stylistic features such as most common words that discriminate each author from another. The idea behind the method is omitting in iteration the most discriminating features gradually one by one, and then degradation in accuracy is measured. So documents that have originated from the same author gain very low accuracy as the discriminative features are removed. Contrarily documents originating from different authors keep the same accuracy without a salient drop. This technique measures the depth differences between two texts from the same author and therefore is assumed to be an endorsement method for the performance of the detection system.

Other future work for extrinsic methods is conducting more investigative steps by applying latent semantic analysis using different corpora. Applying LSA in different corpora is useful to generate different semantic spaces, which are

created by applying different parameters. These parameters include investigating the methods that are related to create the pseudo-documents such as vector-sum or folding-in and compare their performance separately.

Intrinsic methods for plagiarism detection

A number of directions for refining the proposed method have been planned by using different language datasets. In addition the author intends to add a set of least common words to evaluate the performance of the method. Furthermore the author intends to identify specific lists of both least and most common words for future investigations in the area of authorship detection and forensic analysis.

A great variety of machine-learning algorithms, formulas, and techniques exist, however probably the most important aspect of machine-learning methods is the choice of features. The choice of machine learning method, and other specifics of training, is also of great importance to the ultimate success of the classifier. The testing of different machine-learning models, although fairly representative, was not complete. The two main ways in which this exploration of different models could be improved is in trying a greater range of different models, especially meta-models, and in further tweaking the parameters of the machine-learning model (i.e. beyond the features of this study). For the most part, no special refinement of parameters (e.g. forest size in number of trees, for Random Forest) was implemented. More investigation for both approaches by applying different parameters for the used machine learning algorithms will be scheduled.

Stylometric methods have a long history, even if they have only recently begun to receive systematic computational treatment and enhancement. In fact, machine-learning methods represent the pinnacle of current stylometric methodology. The use of statistical properties of words is related to, but distinct in this case, from multivariate approaches to stylometry. The use of relative frequencies constitutes a novel feature type that may find more use in the future, in follow-up experiments. For future plans more advanced stylometric

features could be generated in a number of ways. One possible novel source of features could be to investigate the scores of all books that were correctly classified, both for the actual author class and for all the incorrect author classes. The values of the different parameters for “author” vs “non-author” could be used as features in a secondary machine-learning algorithm, a second layer to process the training output of the initial, LSA algorithm. Thus, training would still use all authors (all classes), but testing would use the secondary layer of machine learning, trained upon the scores of parameters for the algorithm.

7.5 **Summary**

This thesis set out to demonstrate that the existing plagiarism detection systems can be improved beyond the capability of the current approaches. The strong point of this research comes from the fact that the proposed approaches rely on the combination of well-performed methods that demonstrated their high capability in different applications. These methods were used to determine that the plagiarism detection challenges can be addressed when well-performed methods work together. Latent semantic analysis as an intelligent technique has been used in several applications that need to emulate human ability such as essay scoring. It was evidenced that LSA can grade the content of essay exams just as well as people. Another method that called stylometry which approved its ability to quantify the writing style attributes in order to identify text authorship. The advent of machine learning and the high reputation that has been assigned to its algorithms in particular in the field of authorship analysis represents another encouraging factor for methods combination. Further to selecting the aforementioned methods which were considered as suitable for the plagiarism tasks, the selection of most common words frequencies as the main parameter for both approaches has established the baseline for further development to build upon. Two approaches were proposed by this study to address the limitations in extrinsic and intrinsic methods for plagiarism detection. An extrnsic method was proposed for plagiarism detection that was based on the integration between LSA, stylometry and machine learning algorithms. LSA was

fine-tuned to take-in the MCW as stylistic features in order to complement the tasks of semantics detection and capturing authorial attributes to enhance the classification procedure. The results revealed that this approach has scored good classification accuracy. For intrinsic methods for detecting plagiarism, the frequencies of MCW was the pinnacle of this approach, it was used with more than three derived feature sets to train four machine learning algorithms. The model has been tested to verify if the target book was written by the target author or not. LSA was involved in this approach to shrink the vectors space that resulted from bag of words (BOW). BOW is the technique that was used for both approaches to generate the initial features set. However BOW has generated all texts features (words) for the extrinsic method and just the MCW frequencies were generated for the intrinsic method. The results revealed the proposed intrinsic method has scored highly by using one of the proposed feature sets for this research. The most informative feature set were those dealing with standard deviation, especially standard deviation of the in-series proportions of most common words. This represents one of the most original contributions of this work as the relative frequencies of words (as opposed to the raw frequencies) have not yet been reported in the literature.

References

Abbasi, A. and Chen, H., (2008) 'Writeprints: A Stylometric Approach to Identity level Identification and Similarity Detection in Cyberspace'. *ACM Transactions on Information Systems (TOIS)* 26 (2), 7

Afroz, S., (2013) *Deception in Authorship Attribution*. Doctoral dissertation: Drexel University

Al Batineh, M.S. (2015) *Latent Semantic Analysis, Corpus Stylistics and Machine Learning Stylometry for Translational and Authorial Style Analysis: The Case of Denys Johnson-Davies' Translations into English*. Doctoral dissertation: Kent State University

Al Batineh, M.S. (2015) *Latent Semantic Analysis, Corpus stylistics and Machine Learning Stylometry for Translational and Authorial Style Analysis: The Case of Denys Johnson-Davies' Translations into English* (Doctoral dissertation, Kent State University).

Alsallal, M. and Iqbal, R. (2013) An Approach to Detect Illegal Similarity in Research Literature Using Latent Semantic Indexing. In: Plagiarism across Europe and Beyond. [online] available from <*https://plagiarism.pefka.mendelu.cz/files/proceedings.pdf*> [2/11/16]

Alsallal, M., Iqbal, R., Amin, S. and James, A., (2013) 'Intrinsic Plagiarism Detection Using Latent Semantic Indexing And Stylometry'. IEEE. In *Developments in eSystems Engineering (DeSE), 2013 Sixth International Conference on* 145-150

Altınel, B., Ganiz, M.C. and Diri, B. (2015) 'A Corpus-Based Semantic Kernel For Text Classification By Using Meaning Values Of Terms'. *Engineering Applications of Artificial Intelligence* 43, 54 – 66

Alzahrani, S. and Salim, N., (2010) 'Fuzzy semantic-based string similarity for extrinsic plagiarism detection'. *Lab Report for PAN at CLEF 2010 .Braschler and Harman*.

Alzahrani, S. M., Salim, N. and Abraham, A. (2012) 'Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods'. *IEEE Transactions on Systems, Man, and Cybernetics* 42 (2), 133-149

Alzahrani, S.M., Salim, N. and Palade, V. (2015) 'Uncovering Highly Obfuscated Plagiarism Cases Using Fuzzy Semantic-Based Similarity Model'. *Journal of King Saud University-Computer and Information Sciences*, *27* (3), 248-268.

Angoff, W.H. (1974) 'The Development Of Statistical Indices For Detecting Cheaters'. *Journal of the American Statistical Association*, *69* (345), 44-49.

Argamon, S., Bloom, K., Esuli, A. and Sebastiani, F. (2007) 'Automatically Determining Attitude Type and Force for Sentiment Analysis'. In *Language and Technology Conference* (218-231)

Argamon, S., Šarić, M. and Stein, S.S., (2003) 'Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results'. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 475-480). ACM.

Bailey, J.H. and Ottaway, G.H., (1979) International Business Machines Corporation, *Video Recording Disk with Interlacing of Data For Frames on The Same Track*. U.S. Patent 4,161,753.

Baroni, M. and Bernardini, S. (2004) BootCaT: Bootstrapping Corpora and Terms from the Web. In *LREC*.

Berry, J.W., (1992) 'Acculturation and Adaptation In A New Society'. *International migration 30* (1), 69-85

Berry, M.W., Dumais, S.T. and O'Brien, G.W., (1994) 'Using Linear Algebra For Intelligent Information Retrieval'. *SIAM review 37* (4), 573-595

Biber, D., (1995) '*Dimensions of register variation: A cross-linguistic comparison'*. Cambridge: University Press.

Boser, B.E. and Guyon, I.M. (1992) VN Vapnik in 5th Annual ACM Workshop on COLT, D. Haussler

Boser, B.E. and Guyon, I.M., (1992). VN Vapnik in 5th Annual ACM Workshop on COLT, ed by D. Haussler.

Boukhaled, M.A. and Ganascia, J.G., (2015) 'Using Function Words for Authorship Attribution: Bag-Of-Words vs. Sequential Rules'. *Natural Language Processing and Cognitive Science: Proceedings 2014*, 115.

Boulis, C. and Ostendorf, M., (2005) 'Text Classification By Augmenting The Bag-Of-Words Representation with Redundancy-Compensated Bigrams'. *International Workshop in Feature Selection in Data Mining* 9-16

Brin, S., Davis, J. and Garcia-Molina, H., (1995) 'Copy Detection Mechanisms for Digital Documents'. *ACM SIGMOD Record* 24 (2), 398-409

Britt, M.A., Wiemer-Hastings, P., Larson, A.A. and Perfetti, C.A., (2004) 'Using Intelligent Feedback To Improve Sourcing And Integration In Students' Essays'. *International Journal of Artificial Intelligence in Education 14* (3, 4), 359-374

Bull, J., Colins, C., Coughlin, E. and Sharp, D. (2001) Technical review of plagiarism detection software report.

Burges, C.J., (1998) 'A Tutorial On Support Vector Machines For Pattern Recognition'. *Data mining and knowledge discovery* 2 (2), 21-167

Burrows, J., (2002) 'Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and linguistic computing 17* (3), 267-287

Burrows, J.F., (1987) Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and linguistic Computing 2* (2), 61-70

Butler, D., (2010) 'Journals Step Up Plagiarism Policing'. *Nature*, *466* (7303), 167-167

Camargo, L.S. and Yoneyama, T., (2001) 'Specification Of Training Sets and the Number of Hidden Neurons for Multilayer Perceptrons'. *Neural Computation*, *13* (12), 2673-2680

Carroll, J. (2007) *A Handbook for Deterring Plagiarism in Higher Education*. Vol. 2. Oxford: Oxford Centre for Staff and Learning Development.

Ceska, Z., (2009) *Automatic plagiarism detection based on latent semantic analysis*. Ph. D. dissertation, Faculty Appl Sci., Univ. West Bohemia, Pilsen, Czech Republic.

Chaski, C.E., (2005). 'Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigation's'. *International journal of digital evidence*, *4* (1), 1-13

Chester, G., (2001) Pilot of Free-text Electronic Plagiarism Detection Software [online]available from < *http://tinyurl. com/agbezc2* >[1/11/16]

Chong, M.Y.M., (2013) A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques. PhD Thesis: University of Wolverhampton

Chung, C. and Pennebaker, J.W., (2007) 'The Psychological Functions of Function Words'. *Social Communication*, 343-359.

Clough, P. (2003) 'Old and New Challenges in Automatic Plagiarism Detection'. *Plagiarism Advisory Service*, 1-14

Clough, P., (2000) 'Plagiarism In Natural And Programming Languages: An Overview Of Current Tools And Technologies'. *Plagiarism – overview and current tools,* 1-13

Corti, L., Foster, J. and Thompson, P., (1995) 'Archiving Qualitative Research Data'. *Social Research Update*, *10*, 463-91

Cosma, G. (2008) 'An Approach to Source-Code Plagiarism Detection and Investigation Using Latent Semantic Analysis'. PhD Thesis: *University of Warwick, Department of Computer Science*

Cosma, G. and Joy, M., (2012) 'An Approach To Source-Code Plagiarism Detection And Investigation Using Latent Semantic Analysis'. *IEEE transactions on computers 61* (3), 379-394

Coyotl-Morales, R.M., Villaseñor-Pineda, L., Montes-y-Gómez, M. and Rosso, P., (2006) 'Authorship Attribution Using Word Sequences'. *Congress on Pattern Recognition* 844-853

Coyotl-Morales, R.M., Villaseñor-Pineda, L., Montes-y-Gómez, M. and Rosso, P. (2006) Authorship Attribution using Word Sequences. In *Iberoamerican Congress on Pattern Recognition* (844-853)

Craig, H. (2004) *Stylistic Analysis and Authorship Studies*. England: Blackwell Publishing

Culwin, F. and Lancaster, T. (2001) 'Plagiarism issues for higher education'. *Vine 31* (2), 36-41

Culwin, F. and Lancaster, T., (2001) 'Visualising intra-corpal plagiarism'. *Information Visualisation, 2001. Proceedings. Fifth International Conference on* 289-296 IEEE.

De Jager, K. and Brown, C., (2010) 'The Tangled Web: Investigating Academics' views of plagiarism at the University of Cape Town'. *Studies in Higher Education*, *35* (5), 513-528.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990) 'Indexing By Latent Semantic Analysis'. *JASIS,* 41(6), 391-407.

Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003) Authorship Attribution with Support Vector Machines. *Applied intelligence 19* (1-2), 109-123

Diederich, J., Kindermann, J., Leopold, E. and Paass, G., (2000) 'Authorship Attribution With Support Vector Machines'. *Applied intelligence*, *19* (1-2), 109-123.

Dodig-Crnkovic, G., (2002) 'Scientific Methods In Computer Science'. In *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia* (pp. 126-130).

Dumais, S.T. (1991) 'Improving the Retrieval of Information from External Sources' *Behavior Research Methods, Instruments, & Computers*, *23*(2), 229-236

Edmunds, A. and Morris, A., (2000) 'The Problem Of Information Overload In Business Organisations: A Review Of The Literature'. *International journal of information management*, *20* (1), 17-28

Eisa, T.A.E., Salim, N. and Alzahrani, S.,(2015) 'Existing plagiarism detection techniques: A systematic mapping of the scholarly literature'. *Online Information Review*, *39* (3), 383-400

Eiselt, M.P.B.S.A. and Rosso, A.B.C.P., (2009) 'Overview of the 1st international competition on plagiarism detection'. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (p. 1).

Elhadi, M. and Al-Tobi, A., (2008) 'November. Use of text syntactical structures in detection of document duplicates'. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on* (pp. 520-525).

Elio, R., Hoover, J., Nikolaidis, I., Salavatipour, M., Stewart, L. and Wong, K., (2011) About Computing Science Research Methodology. London: Springer

Farringdon, J.M., Morton, A.Q., Farringdon, M.G. and Baker, M.D., (1996) *Analysing for Authorship: A Guide to the Cusum Technique*. University of Wales Press.

Fellbaum, C., (1998) 'A semantic network of English verbs'. *WordNet: An electronic lexical database*, *3*, pp.153-178.

Firth, J.R., (1957) *Papers in linguistics, 1934-1951*. London: Oxford University Press.

Frontini, F., Boukhaled, M.A. and Ganascia, J.G. (2015) Linguistic Pattern Extraction and Analysis for Classic French Plays. In *Presentation at the CONSCILA Workshop, Paris*.

Fucks, W., (1952) 'On Mathematical Analysis of Style'. *Biometrika*, *39*(1/2), pp.122-129

Gamon, M. (2004) 'Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis'. In *Proceedings of the 20th international conference on Computational Linguistics* (841). Association for Computational Linguistics

Ganapathibhotla, M. and Liu, B. (2008) 'Mining opinions in comparative sentences'. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 241-248). Association for Computational Linguistics

Garccıa-Molina, H. and Shivakumar, N., (1995) 'A Copy Detection Mechanism For Digital Documents'. In *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries* (78-89).

Gipp, B. and Meuschke, N., (2011) 'Citation Pattern Matching Algorithms For Citation-Based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking And Longest Common Citation Sequence. In *Proceedings of the 11th ACM symposium on Document engineering* (249-258)

Glendinning, I. (2012) 'European Responses to Student Plagiarism in Higher Education'. In *Proceedings of the 5th International Integrity and Plagiarism Conference*.

Gruner, S. and Naven, S., (2005) 'Tool Support For Plagiarism Detection In Text Documents'. In *Proceedings of the 2005 ACM symposium on Applied computing* (776-781)

Guenther, F.H., Ghosh, S.S. and Tourville, J.A., (2006) 'Neural Modeling And Imaging Of The Cortical Interactions Underlying Syllable Production'. *Brain and language*, *96* (3), 280-301

Hammersley, M., (1997) 'Qualitative Data Archiving: Some Reflections On Its Prospects And Problems'. *Sociology*, *31*(1), 131-142.

Hannabuss, S. (2001) 'Contested Texts: Issues of Plagiarism'. *Library management*, 22(6/7), 311-318

Heaton, J., (2008) 'Secondary Analysis Of Qualitative Data: An Overview'. *Historical Social Research/Historische Sozialforschung*, 33-45

Hollingsworth, C.D., (2012) *Syntactic Stylometry: Using Sentence Structure for Authorship Attribution*

Holmes, D.I., (1998) 'The Evolution Of Stylometry In Humanities Scholarship'. *Literary and linguistic computing 13* (3), 111-117

Honore, L.H., Dill, F.J. and Poland, B.J., (1976) 'Placental Morphology In Spontaneous Human Abortuses With Normal And Abnormal Karyotypes'. *Teratology*, *14*(2), 151-166

Japkowicz, N. and Stephen, S., (2002) 'The class imbalance problem: A systematic study'. *Intelligent data analysis 6* (5), 429-449

Joachims, T., (1998) 'Text categorization with support vector machines: Learning with many relevant features'. *European conference on machine learning* 137-142

Jorge-Botana, G., Leon, J.A., Olmos, R. and Escudero, I. (2010) 'Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora' *Journal of Quantitative Linguistics 17* (1),1-29

Jorge-Botana, G., Leon, J.A., Olmos, R. and Escudero, I., (2010) 'Latent semantic analysis parameters for essay evaluation using small-scale corpora'. *Journal of Quantitative Linguistics 17*(1), 1-29

Juola, P. and Baayen, R.H., (2003) 'A controlled-corpus experiment in authorship identification by cross-entropy'. *Literary and Linguistic Computing*,*20,* 9-67

Kakkonen, T. and Mozgovoy, M. (2010) 'Hermetic and Web Plagiarism Detection Systems for Student Essays: An Evaluation of the State-of-the-Art'. *Journal of Educational Computing Research* 42 (2), 135-159

Köhler, K., Weber-Wulff, D., Barron-Cedeno, A. and Rosso, P., (2010) Plagiarism Detection Test 2010

Kon, M.A. and Plaskota, L., (2000) 'Information Complexity Of Neural Networks'. *Neural Networks*, *13*(3), 365-375

Kontostathis, A. and Pottenger, W.M. (2006) 'A Framework for Understanding Latent Semantic Indexing (LSI) Performance' *Information Processing & Management 42*(1), 56-73

Koppel, M. and Schler, J., (2003) 'Exploiting Stylistic Idiosyncrasies For Authorship Attribution'. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (69), 72

Koppel, M. and Schler, J., (2004) 'Authorship Verification As A One-Class Classification Problem'. In *Proceedings of the twenty-first international conference on Machine learning* 62

Koppel, M., Argamon, S. and Shimoni, A.R., (2002) 'Automatically Categorizing Written Texts by Author Gender'. *Literary and Linguistic Computing 17*(4), 401-412

Koppel, M., Schler, J. and Argamon, S., (2009) 'Computational Methods In Authorship Attribution'. *Journal of the American Society for information Science and Technology 60* (1), 9-26

Koppel, M., Schler, J. and Zigdon, K, (2005) 'Determining An Author's Native Language By Mining A Text For Errors'. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* 624-628

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., (2007) Supervised Machine Learning: A Review Of Classification Techniques

Koch, M.W., Moya, M.M., Hostetler, L.D. and Fogler, R.J., 1995. 'Cueing, feature discovery, and one-class learning for synthetic aperture radar automatic target recognition'. *Neural Networks*, *8*(7), pp.1081-1102.

Lahiri, S., & Mihalcea, R. (2013) 'Using N-Gram And Word Network Features For Native Language Identification'. *NAACL/HLT 2013*, 251

Lampson, B., Abadi, M., Burrows, M. and Wobber, E., (1992) 'Authentication in Distributed Systems: Theory and Practice'. *ACM Transactions on Computer Systems (TOCS)*, *10*(4), 265-310

Lancaster, T. and Culwin, F., (2004) 'A Visual Argument For Plagiarism Detection Using Word Pairs'. In *Plagiarism: Prevention, Practice and Policy Conference*.

Lancaster, T. and Culwin, F., (2005) 'Classifications Of Plagiarism Detection Engines'. *Innovation in Teaching and Learning in Information and Computer Sciences*, *4*(2), 1-16

Landauer, T. K., & Dumais, S. T. (1997) 'A Solution To Plato's Problem: The Latent Semantic Analysis Theory Of The Acquisition, Induction, And Representation Of Knowledge. *Psychological Review,* 104, 211-240

Landauer, T.K., Foltz, P.W. and Laham, D., (1998) 'An introduction to latent semantic analysis'. *Discourse processes*, *25*(2-3), 259-284

Lane, P.C., Lyon, C. and Malcolm, J.A., (2006) 'Demonstration Of The Ferret Plagiarism Detector'. *Proceedings of the 2nd International Plagiarism Conference*.

Lanthrop, A. and Foss, K., (2000) Student Cheating And Plagiarism In The Internet Era.

Larkham, P.J. and Manns, S. (2002) 'Plagiarism and its Treatment in Higher Education'. *Journal of Further and Higher Education* 26 (4), 339-349.

Liu, H. and Motoda, H., (2001) 'Data reduction via instance selection'. *Instance selection and construction for data mining* 3-20

Louw, B., (1993) 'Irony In The Text Or Insincerity In The Writer? The Diagnostic Potential Of Semantic Prosodies'. *Text and technology: In honour of John Sinclair*, *157*, 176

Lukashenko, R., Graudina, V. and Grundspenkis, J. (2007) 'Computer-Based Plagiarism Detection Methods and Tools: An Overview' In *Proceedings of the 2007 international conference on Computer systems and technologies* (p. 40). ACM.

Luyckx, K. and Daelemans, W. (2008) 'Authorship Attribution and Verification with Many Authors and Limited Data' In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 513-520). Association for Computational Linguistics

Lyon, C., Malcolm, J. and Dickerson, B., (2001) 'Detecting Short Passages Of Similar Text In Large Document Collections'. In *Proceedings of the 2001 conference on empirical methods in natural language processing* 118-125

Madigan, D., Einahrawy, E., Martin, R.P., Ju, W.H., Krishnan, P. and Krishnakumar, A.S. (2005) 'Bayesian Indoor Positioning Systems'. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies* 2, 1217-1227

Markovitch, S. and Rosenstein, D., (2002) 'Feature Generation Using General Constructor Functions'. *Machine Learning 49*(1), 59-98

Marsh, B., (2004) 'Turnitin. Com and the Scriptural Enterprise Of Plagiarism Detection'. *Computers and Composition 21*(4), 427-438

Matthews, R.A. and Merriam, T.V., (1994) 'Neural Computation In Stylometry I: An Application To The Works Of Shakespeare And Fletcher'. *Literary and Linguistic Computing 8* (4), 203-209

Maurer, H. and Zaka, B., (2007) 'Plagiarism–A Problem And How To Fight It'. *Proceeding of Ed-Media 2007*, 4451-4458

Maurer, H.A., Kappe, F. and Zaka, B., (2006) 'Plagiarism-A Survey'. *J. UCS*, *12* (8),1050-1084

McCarthy, P.M., Lewis, G.A., Dufty, D.F. and McNamara, D.S., (2006) Analyzing Writing Styles with Coh-Metrix. In *FLAIRS Conference* 764-769

Mendenhall, T.C., (1887) 'The Characteristic Curves Of Composition'. *Science*, 237-249

Monostori, K., Zaslavsky, A. and Schmidt, H., (2000) 'Document Overlap Detection System For Distributed Digital Libraries'. In *Proceedings of the fifth ACM conference on Digital libraries* 226-227

Montes-y-Gómez, M., Villasenor-Pineda, L. and Rosso, P., (2013) 'Determining and Characterizing the Reused Text for Plagiarism Detection'. *Expert Systems with Applications* 40 (5), 1804-1813

Moore, A.F., (2001) *Rock, The Primary Text: Developing A Musicology Of Rock*. Ashgate Pub Ltd.

Mosteller, F. and Wallace, D., (1964) 'Inference And Disputed Authorship: *The Federalist'*

Mozgovoy, M., Kakkonen, T. and Cosma, G. (2010) 'Automatic Student Plagiarism Detection: Future Perspectives'. *Journal of Educational Computing Research* 43 (4), 511-531.

Murray, L.J. (2006) 'Plagiarism and Copyright Infringement: The Costs of Confusion'. In *Originality, Imitation, and Plagiarism.* Ed, by Eisner, C. & Vicinus, M. Canada: Queen's University, 1-22

Myers, S. (1998) 'Questioning Author (ity): ESL/EFL, Science, and Teaching about Plagiarism'. *TESL-EJ* 3 (2), 1-15.

Nadeau, C. and Bengio, Y., (2003) 'Inference for the Generalization Error'. *Machine Learning*, *52*(3), 239-281

Neocleous, C. and Schizas, C., (2002) 'Artificial Neural Network Learning: A Comparative Review'. *Hellenic Conference on Artificial Intelligence* 300-313

Oakes, M.P. and Ji, M. eds., (2012) '*Quantitative Methods In Corpus-Based Translation Studies: A Practical Guide To Descriptive Translation Research* (Vol. 51)'. New York: John Benjamins Publishing.

Oberreuter, G., L'Huillier, G., Rıos, S.A. and Velásquez, J.D. (2011) 'Approaches For Intrinsic And External Plagiarism Detection'. *Proceedings of the PAN* at CLEF 2011.Chile: University of Chile

Ottenstein, K.J. (1976) 'An Algorithmic Approach to the Detection and Prevention of Plagiarism' *ACM Sigcse Bulletin* 8(4), 30-41

Ottenstein, K.J., (1976) 'An Algorithmic Approach To The Detection And Prevention Of Plagiarism.' *ACM Sigcse Bulletin*, *8*(4), 30-41

Peng, F., Schuurmans, D. and Wang, S. (2004) 'Augmenting Naive Bayes Classifiers with Statistical Language Models' *Information Retrieval 7* (3-4), 317-345

Piao, S.S., Clough, P., and Arundel, J., (2001) 'Proposing Basic Approaches to Detecting Text Rewrite'. University of Sheffield

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2009). 'Cross-Language Plagiarism Detection'. *Language Resources and Evaluation*, *45*(1), 45-62

Purdy, J.P. and Walker, J.R., (2013) 'Liminal Spaces And Research Identity The Construction Of Introductory Composition Students As Researchers'. *Pedagogy 13*(1), 9-41

Quinlan, J.R., (1986) 'Induction Of Decision Trees'. *Machine learning 1*(1), 81-106

Ramnial, H., Panchoo, S. and Pudaruth, S. (2016) 'Authorship Attribution Using Stylometry and Machine Learning Techniques'. *Intelligent Systems Technologies and Applications* 113-125

Redfern, K. and Barnwell, N. (2009) 'Cultural Differences in Attitudes toward Plagiarism in Undergraduate Business Students': An Empirical Investigation. Unpublished Thesis: University of Technology Sydney Australia

Řehůřek, R., (2007) 'Subspace Tracking for Latent Semantic Analysis'. In *European Conference on Information Retrieval* 289-300

ŘEHŮŘEK, R., (2008) *Semantic-based plagiarism detection* (Doctoral dissertation: Masarykova univerzita, Fakulta informatiky

Roig, M. (2001) 'Plagiarism and Paraphrasing Criteria of College and University Professors'. *Ethics & Behaviour* 11 (3), 307 -323

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., (1985) '*Learning Internal Representations By Error Propagation'* (No. ICS-8506). CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Roberts, S. and Tarassenko, L., 1994. A probabilistic resource allocating network for novelty detection. *Neural Computation*, *6*(2), pp.270-284.

Salton, G., Wong, A. and Yang, C.S., (1975) 'A vector space model for automatic'

Sánchez-Vega, F., Villatoro-Tello, E., Montes-y-Gómez, M., Villasenor-Pineda, L. and Rosso, P. (2013) 'Determining and Characterizing the Reused Text for Plagiarism Detection' *Expert Systems with Applications 40* (5), 1804-1813

Sanderson, C. and Guenter, S., (2006) 'Short Text Authorship Attribution via Sequence Kernels, Markov Chains, And Author unmasking: An Investigation'. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (482-491). Association for Computational Linguistics

Sebastiani, F. (2002) 'Machine Learning in Automated Text Categorization' *ACM Computing surveys (CSUR) 34*(1), 1-47

Shivakumar, N. and Garcia-Molina, H., (1996) 'Building A Scalable And Accurate Copy Detection Mechanism'. In *Proceedings of the first ACM international conference on Digital libraries* 160-168

Siddique, M.N.H. and Tokhi, M.O. (2001) 'Training Neural Networks: Backpropagation Vs. Genetic Algorithms'. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on* 4, 2673-2678

Sinclair, J., (1991) *Corpus, Concordance, Collocation*. UK: Oxford University Press

Sorokina, D., Gehrke, J., Warner, S. and Ginsparg, P. (2006) 'Plagiarism Detection in arXiv'. IEEE, In *Sixth International Conference on Data Mining (ICDM'06)* 1070-1075

Sousa Silva, R., Grant, T. and Maia, B., (2010) "I didn't mean to steal someone else's words!": a forensic linguistic approach to detecting intentional plagiarism. In *4th International Plagiarism Conference*.

Stamatatos, E., (2009) 'A Survey Of Modern Authorship Attribution Methods'. *Journal of the American Society for information Science and Technology 60* (3), 538-556

Stamatatos, E., (2009) 'A Survey of Modern Authorship Attribution Methods'. *Journal of the American Society for information Science and Technology 60* (3), 538-556

Stamatatos, E., Fakotakis, N. and Kokkinakis, G., (2000) 'Automatic Text Categorization In Terms Of Genre And Author'. *Computational linguistics 26* (4), 471-495

Stappenbelt, B. and Rowles, C. (2010) 'The Effectiveness of Plagiarism Detection Software as Learning Tool in Academic Writing Education'. In *4th Asia Pacific Conference on Educational Integrity (4APCEI)* 29

Stein, B. and zu Eissen, S.M., (2007) Intrinsic Plagiarism Analysis with Meta Learning. *PAN*, *276*.

Stein, B., Lipka, N. and Prettenhofer, P., (2011) 'Intrinsic Plagiarism Analysis'. *Language Resources and Evaluation*, *45*(1), 63-82

Smith, T.C. and Witten, I.H., 1993. 'Language inference from function words'.

Tashakkori, A. and Teddlie, C., (1998) *Mixed Methodology: Combining Qualitative And Quantitative Approaches*. Vol. 46. UK: Sage

Tax, D.M. and Duin, R.P., 2001. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, *2*(Dec), pp.155-173.

Tesar, R., Strnad, V., Jezek, K. and Poesio, M., (2006) 'Extending The Single Words-Based Document Model: A Comparison Of Bigrams And 2-Itemsets'. In *Proceedings of the 2006 ACM symposium on Document engineering* 138-146

Tsatsaronis, G., Varlamis, I., Giannakoulopoulos, A. and Kanellopoulos, N., (2010) 'Identifying Free Text Plagiarism Based On Semantic Similarity'. In *Proceedings of the 4th International Plagiarism Conference*

Turney, P.D. and Pantel, P., (2010) 'From Frequency To Meaning: Vector Space Models Of Semantics'. *Journal of artificial intelligence research 37* (1),141-188

Tweedie, F.J., Singh, S. and Holmes, D.I., (1996) 'Neural Network Applications In Stylometry: The Federalist Papers'. *Computers and the Humanities*, *30* (1), 1-10

Uzuner, Ö. and Katz, B., (2005) 'A Comparative Study Of Language Models For Book And Author Recognition'. In *International Conference on Natural Language Processing* 969-980

Wallach, H.M., (2006) 'Topic Modelling: Beyond Bag-of-Words'. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). ACM.

Weber-Wulff, D., (2008) On The Utility Of Plagiarism Detection Software. In *Third International Plagiarism Conference, Gateshead.*

Weigend, A.S., Rumelhart, D.E. and Huberman, B.A., (1991) Generalization By Weight-Elimination Applied To Currency Exchange Rate Prediction. In *Neural*

Networks, 1991., IJCNN-91-Seattle International Joint Conference on (1) 837-841

White, D.R. and Joy, M.S., (2004) 'Sentence-Based Natural Language Plagiarism Detection'. *Journal on Educational Resources in Computing (JERIC) 4* (4), 2

Williams, J.B. (2002) 'The Plagiarism Problem: Are Students Entirely to Blame' In *Proceedings 19th ASCILITE Conference*

Wise, M.J., (1996) 'YAP3: Improved Detection Of Similarities In Computer Program And Other Texts'. *ACM SIGCSE Bulletin 28* (1), 130-134

Yam, J.Y. and Chow, T.W., (2001) 'Feedforward Networks Training Speed Enhancement By Optimal Initialization Of The Synaptic Coefficients'. *IEEE Transactions on Neural Netw 12* (2), 430-434

Yen, G.G. and Lu, H., (2000) 'Hierarchical Genetic Algorithm Based Neural Network Design'. In *Combinations of Evolutionary Computation and Neural Networks, 2000 IEEE Symposium on* 168-175

Yerra, R. and Ng, Y.K., (2005) 'A Sentence-Based Copy Detection Approach For Web Documents'. *International Conference on Fuzzy Systems and Knowledge Discovery* 557-570

Yu, L. and Liu, H., (2004) 'Efficient Feature Selection Via Analysis Of Relevance And Redundancy'. *Journal of machine learning research 5*, 1205-1224

Yule, G.U., (1944) 'On Sentence-Length As A Statistical Characteristic Of Style In Prose: With Application To Two Cases Of Disputed Authorship'. *Biometrika 30*(3/4), 363-390.

Zechner, M., Muhr, M., Kern, R., & Granitzer, M. (2009) 'External and Intrinsic Plagiarism Detection using vector Space Models'. In Proc. of 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection 47-55

Zelenko, L., (2003) 'Construction of the Essential Spectrum for a Multidimensional Non-self-adjoint Schrödinger Operator via the Spectra of Operators with Periodic Potentials', I. *Integral Equations and Operator Theory*, *46*(1), pp.11-68.

Zhang, G.P., (2000) 'Neural Networks for Classification: A Survey'. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(4), pp.451-462.

Zhang, H.Y., (2010) 'CrossCheck: an effective tool for detecting plagiarism'. *Learned Publishing*, *23*(1), pp.9-14.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758-2765.

Zhao, Y. and Zobel, J., (2005) 'Searching With Style: Authorship Attribution In Classic Literature'. In *Proceedings of the thirtieth Australasian conference on Computer science 62*, 59-68

Zheng, R., Li, J., Chen, H. and Huang, Z., (2006) 'A Framework for Authorship Identification Of Online Messages: Writing-Style Features And Classification Techniques'. *Journal of the American Society for Information Science and Technology*, *57* (3), 378-393.

Zu Eissen, S. M., & Stein, B. (2006) 'Intrinsic Plagiarism Detection'. *European Conference on Information Retrieval* 565-569

Zu Eissen, S.M., Stein, B. and Kulig, M. (2007) 'Plagiarism Detection without Reference Collections'. *Advances in data analysis* 359-366

Zurini, M. (2015) 'Stylometry Metrics Selection for Creating a Model for Evaluating the Writing Style of Authors According to Their Cultural Orientation'. *Informatica Economica* 19 (3), 107

Appendix A

List of publications

Alsallal, M. and Iqbal, R. (2013) An Approach to Detect Illegal Similarity in Research Literature Using Latent Semantic Indexing. In: Plagiarism across Europe and Beyond. [online] available from <*https://plagiarism.pefka.mendelu.cz/files/proceedings.pdf*> [2/11/16]

Alsallal, M., Iqbal, R., Amin, S. and James, A., (2013) 'Intrinsic Plagiarism Detection Using Latent Semantic Indexing And Stylometry'. IEEE. In *Developments in eSystems Engineering (DeSE), 2013 Sixth International Conference on* 145-150.

Alsallal, M., Iqbal, R., Amin, S. James, A., and Palade V., (2016) 'I An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection'. IEEE. In *Developments in Developments of eSystems Engineering (DeSE), 2016 9th International Conference*.

Publication in Progress

Journal: An Integrated Approach for Intrinsic Plagiarism Detection: Muna Al-Sallal, Rahat Iqbal, Victor Chang and Vasile Palade. This paper has been accepted by Future Generation Computer Systems Journal.

Journal: Ltent Semantic Analysis based Setylometry for extrinsic Plagiarism Detection, to be submitted to Artificial Intelligence Journal.

Appendix B

Low Risk Research Ethics Approval Checklist

Chapter 8pplicant Details

Name: Muna Alsallal	**E-mail:** **alsallam@uni.coventry.ac.uk**
Department: Computing	**Date:**
Course: Computing	**Title of Project:** A Machine Learning Approach for Plagiarism Detection

Chapter 9Project Details

This thesis has developed two novel approaches to address the extrinsic and intrinsic methods. Firstly a novel extrinsic method for detecting plagiarism is proposed. The method is based on four well-known techniques namely Bag of Words (BOW), Latent Semantic Analysis (LSA), Stylometry and machine learning. The LSA application was fine-tuned to take in the stylometric features (most common words) in order to characterise the document authorship

Participants in your research

Will the project involve human participants?	Yes	No ✓

If you answered **Yes** to this questions, this may **not** be a low risk project.

If you are a student, please discuss your project with your Supervisor.

If you are a member of staff, please discuss your project with your Faculty Research Ethics Leader or use the Medium to High Risk Ethical Approval or NHS or Medical Approval Routes.

Risk to Participants

Will the project involve human patients/clients, health professionals, and/or patient (client) data and/or health professional data?	Yes	No ✓
Will any invasive physical procedure, including collecting tissue or other samples, be used in the research?	Yes	No ✓
Is there a risk of physical discomfort to those taking part?	Yes	No ✓
Is there a risk of psychological or emotional distress to those taking part?	Yes	No ✓
Is there a risk of challenging the deeply held beliefs of those taking part?	Yes	No ✓
Is there a risk that previous, current or proposed criminal or illegal acts will be revealed by those taking part?	Yes	No ✓
Will the project involve giving any form of professional, medical or legal advice, either directly or indirectly to those taking part?	Yes	No ✓

If you answered **Yes** to **any** of these questions, this may **not** be a low risk project.

If you are a student, please discuss your project with your Supervisor.

If you are a member of staff, please discuss your project with your Faculty Research Ethics Leader or use the Medium to High Risk Ethical Approval or NHS or Medical Approval Routes.

Risk to Researcher

Will this project put you or others at risk of physical harm, injury or death?	Yes	No ✓
Will project put you or others at risk of abduction, physical, mental or sexual abuse?	Yes	No ✓
Will this project involve participating in acts that may cause psychological or emotional distress to you or to others?	Yes	No ✓
Will this project involve observing acts which may cause psychological or emotional distress to you or to others?	Yes	No ✓
Will this project involve reading about, listening to or viewing materials that may cause psychological or emotional distress to you or to others?	Yes	No ✓
Will this project involve you disclosing personal data to the participants other than your name and the University as your contact and e-mail address?	Yes	No ✓
Will this project involve you in unsupervised private discussion with people who are not already known to you?	Yes	No ✓
Will this project potentially place you in the situation where you may receive unwelcome media attention?	Yes	No ✓
Could the topic or results of this project be seen as illegal or attract the attention of the security services or other agencies?	Yes	No ✓
Could the topic or results of this project be viewed as controversial by anyone?	Yes	No ✓

If you answered **Yes** to **any** of these questions, this is **not** a low risk project. Please:

If you are a student, discuss your project with your Supervisor.

If you are a member of staff, discuss your project with your Faculty Research Ethics Leader or use the Medium to High Risk Ethical Approval route.

Informed Consent of the Participant

Are any of the participants under the age of 18?	Yes	No ✓
Are any of the participants unable mentally or physically to give consent?	Yes	No ✓
Do you intend to observe the activities of individuals or groups without their knowledge and/or informed consent from each participant (or from his or her parent or guardian)?	Yes	No ✓

If you answered **Yes** to **any** of these questions, this may **not** be a low risk project. Please:

If you are a student, discuss your project with your Supervisor.

If you are a member of staff, discuss your project with your Faculty Research Ethics Leader or use the Medium to High Risk Ethical Approval route.

Participant Confidentiality and Data Protection

Will the project involve collecting data and information from human participants who will be identifiable in the final report?	Yes	No ✓
Will information not already in the public domain about specific individuals or institutions be identifiable through data published or otherwise made available?	Yes	No ✓
Do you intend to record, photograph or film individuals or groups without their knowledge or informed consent?	Yes	No ✓

Do you intend to use the confidential information, knowledge or trade secrets gathered for any purpose other than this research project?	Yes	No ✓

If you answered **Yes** to **any** of these questions, this may **not** be a low risk project:

If you are a student, discuss your project with your Supervisor.

If you are a member of staff, discuss your project with your Faculty Research Ethics Leader or use the Medium to High Risk Ethical Approval or NHS or Medical Approval routes.

Gatekeeper Risk

Will this project involve collecting data outside University buildings?	Yes	No ✓
Do you intend to collect data in shopping centres or other public places?	Yes	No ✓
Do you intend to gather data within nurseries, schools or colleges?	Yes	No ✓
Do you intend to gather data within National Health Service premises?	Yes	No ✓

If you answered **Yes** to **any** of these questions, this is **not** a low risk project. Please:

If you are a student, discuss your project with your Supervisor.

If you are a member of staff, discuss your project with your Faculty Research Ethics Leader or use the Medium to High Risk Ethical Approval or NHS or Medical Approval routes.

Other Ethical Issues

Is there any other risk or issue not covered above that may pose a risk to you or any of the participants?	Yes	No ✓
Will any activity associated with this project put you or the participants at an ethical, moral or legal risk?	Yes	No ✓

If you answered **Yes** to these questions, this may **not** be a low risk project. Please:

If you are a student, discuss your project with your Supervisor.

If you are a member of staff, discuss your project with your Faculty Research Ethics Leader.

Principal Investigator Certification

If you answered **No** to **all** of the above questions, then you have described a low risk project. Please complete the following declaration to certify your project and keep a copy for your record as you may be asked for this at any time.

Agreed restrictions to project to allow Principal Investigator Certification

Please identify any restrictions to the project, agreed with your Supervisor or Faculty Research Ethics Leader to allow you to sign the Principal Investigator Certification declaration.

Participant Information Leaflet attached. Informed Consent Forms attached. Risk Assessment Form attached.

Principal Investigator's Declaration

Please ensure that you:

Tick all the boxes below and sign this checklist.

Students must get their Supervisor to countersign this declaration.

I believe that this project **does not require research ethics approval**. I have completed the checklist and kept a copy for my own records. I realise I may be asked to provide a copy of this checklist at any time.	✓
I confirm that I have answered all relevant questions in this checklist honestly.	✓
I confirm that I will carry out the project in the ways described in this	✓

checklist. I will immediately suspend research and request a new ethical approval if the project subsequently changes the information I have given in this checklist.	

Signatures

If you or your supervisor do not have electronic signatures, please type your name in the signature space. An email sent from the Supervisor's University inbox will be accepted as having been signed electronically.

Principal Investigator

Signed Muna Alsallal

Date: 30/09/2015

Students storing this checklist electronically must append to it an email from your Supervisor confirming that they are prepared to make the declaration above and to countersign this checklist. This-email will be taken as an electronic countersignature.

Student's Supervisor Rahat Iqbal

Date 30/09/2015

I have read this checklist and confirm that it covers all the ethical issues raised by this project fully and frankly. I also confirm that these issues have been discussed with the student and will continue to be reviewed in the course of supervision.